IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 19, 2018, accepted January 4, 2019, date of publication January 11, 2019, date of current version February 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2892466

Joint Resource Allocation and User Association for Heterogeneous Services in Multi-Access Edge Computing Networks

JIZHE ZHOU, XING ZHANG, AND WENBO WANG

School of Information and Communications Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Xing Zhang (hszhang@bupt.edu.cn)

This work was supported by the National Science Foundation of China under Grant 61771065, Grant 61571054, and Grant 61631005.

ABSTRACT Multi-access edge computing (MEC) has emerged as a promising technique for low-latency services, in light of its proximity to users and embedded cloud computing capability. In order to improve the network efficiency and fairness, it is crucial and challenging to jointly optimize resource management and user association mechanism with the consideration of heterogeneous services and network conditions. Moreover, the existing studies always neglect the heterogeneities of services on the requirements of both computation capability and storage capability. To solve this issue, we derive a strategy to improve the overall delay-aware performance of heterogeneous services with the MEC capability of computation and storage and the choices of the users' association. Accordingly, a coalition-game-based algorithm is proposed to form user coalitions for association scheme and resource sharing policy. In particular, we mathematically present that the proposed algorithm is capable of convergence and optimality. The simulation results also show that our algorithm gets a good performance efficiently. Furthermore, it reduces the weighted sum of delays of users by average 27.8% and 82.1%, while continuously improving delay-aware fairness, compared with those of the priority-based assignment scheme and the nearest assignment scheme, respectively.

INDEX TERMS Multi-access edge computing, resource allocation, user association, coalition game.

I. INTRODUCTION

In the evolvement and advancement to 5G, the prevalence of mobile devices promotes the growth of innovative applications and emerging requirements of services. Concerned with the demands of high computation capability as well as the high energy consumption of devices, cloud computing is envisioned as an effective technology in the past years, which is a central platform of resource management and network provisioning in the core network [1]. The pervasive construction of cloud computing not only support centralized control for resources and networking, but also exposes the problem of bursty traffic burdens on backhaul connections. On the other hand, more heterogeneous services, such as Internet of Things, augmented reality and multimedia, get prevalent for the variety of users' requirement [2]. These services have different requirements for network provisioning and Quality of Service (QoS), which may worsen 5G network degradation as for the explosive traffic data and intensive latency constraint [3], [4].

In light of the urgent challenges, Multi-access Edge Computing (MEC), outlined by ETSI [5], emerges as a representative paradigm for high data rate and low latency services. MEC is expected to provide a seamless cloud-to-thing continuum with the high computation capability and large storage capability in proximity to users. It greatly reduces the distance for transmission and focuses essentially on local services to cater to high Quality of Experience (QoE) requirements for users. By leveraging resources effectively, the MEC-enabled network is a key structure for the prosperity of future application and the alleviation of network pressure to the core network [6], [7].

A. RELATED WORK

As one of the key technologies for future networks, MEC has attracted many attentions in academic and industry fields recently. Computing offloading is viewed as an effective way with the respect of battery energy limitation of mobile devices and latency requirement of services. To balance both network efficiency and user demand, different offloading mechanisms have been proposed in literature. For instance, in [8], a partial offloading scheme is proposed with time scheduling and channel allocation for energy efficiency. Muñoz *et al.* [9]

analyze the tradeoff between energy consumption and latency for computing offloading, and then propose an offloading scheme for partial load offloading. In [10], to minimize the energy consumption of task offloading, an offloading decision is proposed for different applications in the multi-cell network. Some works also investigate computing resource allocation for MEC offloading issue. In [11], in the case of file compression, a joint resource allocation and partial offloading scheme is investigated in terms of the QoE perspective. In [12], considering computing resource constraint in MEC server, a load balancing mechanism is proposed for computing offloading to minimize the time consumption. Moreover, a multitude offloading architecture of MEC and cloud is designed for the delay requirement with resource provisioning and scheduling in [13] and [14], respectively.

On the other hand, caching in edge network is an imperative technology for future network revolution. The benefits of edge caching on QoE improvement and energy saving have been widely studied in past works. Based on video resolution and user demand, a proactive caching model is proposed with QoE provisioning and network status for delay minimization in [15]. Reference [16] also considers multiple video content characteristics and develops a corresponding optimal caching policy. In [17], a learning-based method achieves better prediction on caching decision while reducing network cost. Moreover, the multitude architecture of caching is proved as an influential way for network efficiency in accordance with users' behavior [18] and social relationships [19]. Peng *et al.* [20] also give evidence on the relationship between caching decision and network status, particularly network backhaul capability and edge caching size.

As heterogeneous services process simultaneously in MEC, it is necessary to integrate the distinct requirements for resources by multi-services and collaboratively manage resources with a cooperative purpose. Even though in [2] and [21], computing tasks and data requesting tasks are discussed together, the competition for MEC resources between different services is not clarified.

B. CONTRIBUTIONS

Similar with works in [2] and [21], two main heterogeneous services are considered working in process in a multi-cell network, that is, computing service and content delivery service. MEC is deployed in all small cells. In particular, computing service users intend to offload tasks to MEC with the respect of high computation capability, while content requested of content delivery service can be cached in MEC for short downloading delay. Users of both of the services take the storage space in MEC, while computing service users also make use of the computation resources. Therefore, there exists a problem to distribute resources to different services in the system. Moreover, in traditional work, users are associated with the nearest access points (APs), who may face the problem of inadequate resources in the overloaded MEC. Thus, a flexible user association scheme is necessary with densely deployed APs.

In this paper, a coalition game based algorithm is derived to optimize the delay-aware performance with the consideration of resource allocation and user association scheme. Coalition game is a mathematical tool utilized to predict the strategy of players with group interaction. Past works on resource allocation in wireless communication network show the effectiveness and efficiency of coalition game theory [22]–[24]. In this paper, coalition formation game is applied to solve the problem of resource management and users association in a multi-cell network, and a coalition structure that maximizes the network utility is exploited. Our contributions are summarized as follows.

- A strategy of resource allocation and users association is proposed to minimize the weighted sum of delays of the collaborative heterogeneous services in the MEC network, i.e. computing service and content delivery service. By taking the advantages of edge network, both storage and computation resources are modeled for allocation with the delay-aware network objective.
- The joint resource allocation and users association issue for heterogeneous services is formulated as a mixed integer nonlinear problem (MINLP). Then, we design a coalition game based algorithm to optimize the solution of resource allocation and users association. Theoretical analysis shows that our proposed algorithm can converge to a Nash-stable solution and achieve optimality.
- Finally, simulation results show that our proposed algorithm can reduce the weighted sum of delays compared with the other schemes. The proposed algorithm also outperforms in terms of users' fairness. Moreover, it is presented that our proposed algorithm gets good performance much efficiently than exhaustive search, which illustrates the practicality of the proposed algorithm.

C. ORGANIZATION

The rest of the paper is organized as follows. The system model and the problem formulation are presented in Section II. In Section III, a coalition game is introduced and a coalition based algorithm is proposed. Then, we give theoretical analysis on stability, optimality and convergence in Section IV. The performance of the proposed algorithm is shown in Section V. Finally, conclusions are given in Section VI.

II. SYSTEM OVERVIEW AND MODELS

In this section, we first deliver a system overview of the game-theoretic heterogeneous service collaboration in multi-cell network. Then, the problem of resource allocation and user association in MEC is formulated for further investigation.

A. SYSTEM OVERVIEW

The network framework of heterogeneous services with MEC is presented in Fig.1. There exist one macro base station (MBS) and a set of APs, denoted as $\mathcal{K} = \{1, 2, ..., K\}$, in the network. MBS is the main controller of resource allocation and association decision between users and APs, which is

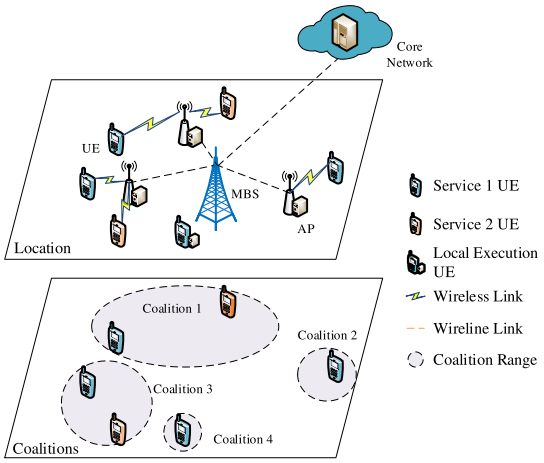

FIGURE 1. A multi-cell MEC network for heterogeneous service users. Coalitions are formed with the collaboration among users.

accessed to core network by backhaul link and connected with APs by wireline link. Each AP is equipped with a MEC server with computation and storage capability. It is assumed that the computation capability of MEC is denoted as F_m, measured in the number of CPU cycles per second [25], [26] and the storage capability is denoted as M bits. It is assumed that bandwidth resources are orthogonal among APs. The overall uplink and downlink bandwidth of a AP are denoted as B^u and B^d, respectively. In this paper, we only analyze the intra-cell interference by the group of associated users. Fig.1 also depicts the coalitions formed by UEs with different APs, which we describe in details in the Section III.

In this paper, user equipments (UEs) are classified into two groups according to the two main heterogeneous services we investigate, i.e., computing service and content delivery service. UEs who requests computing service, denoted as Service 1 UEs in Fig.1, initialize computing tasks which can utilize device processor for computing. Otherwise, users can transmit computing tasks to APs in proximity for processing. On the other hand, the group of content requesting UEs, denoted as Service 2 UEs in Fig.1, derive content downloading requests from the core network server. If the desirable files are previously cached in MEC, UEs can downloading the file directly from MEC. Otherwise, they get the files from the core database. For simplicity, it is assumed that content requesting UEs request independent files. The two groups of UEs mentioned above are denoted as $\mathcal{N}_0 = \{1, 2, ..., |\mathcal{N}_0|\}$ and $\mathcal{N}_1 = \{1, 2, ..., |\mathcal{N}_1|\}$, respectively. $\mathcal{N} = \{\mathcal{N}_c | c = 0, 1\}$ denotes the overall group of UEs in the network, where $c = \{0, 1\}$ denotes the classifications of the two heterogeneous services and $|\mathcal{N}|$ denotes the number of UEs of services.

B. OFFLOADING MODEL

The computing task from a computing service UE $i \in \mathcal{N}_0$ is described as $x_i^0 = \{b_i^0, w_i^0\}$, where b_i^0 denotes the input size of the computing task and w_i^0 denotes the requirement of computing capability measured in CPU cycles [25], [26]. For the perspective of offloading strategy, the offloading decision for UE i is denoted as λ_{ik}^0. Here, $\lambda_{ik}^0 = 1$ if the computing task of UE i is offloaded to the MEC at AP k and 0 otherwise. It is assumed that each UE can only offload task to one AP. It is obviously that $1 - \sum_{k\in\mathcal{K}} \lambda_{ik}^0 = 1$ illustrates that the computing task is executed locally at device.

Then, we discuss the completion delay for computing service. For computing task executed locally, the completion delay is the same as computing delay, expressed as $t_i^0 = \frac{w_i^o}{F_l}$, where F_l denotes the computing capability of local device (in CPU cycles per second). On the other side, if $\lambda_{ik}^0 = 1$, the completion delay consists of uplink transmission delay and computing delay. The uplink transmission rate is denoted as

$$r_{ik}^u = B^u \log_2 \left(1 + \frac{p_i^u G_{ik}^u}{\sum\limits_{j\in\{Z_k^0 \setminus i\}} p_j^u G_{jk}^u + \sigma^2} \right), \tag{1}$$

where p_i^u is the uplink transmission power, G_{ik}^u is the channel gain from UE i to AP k and σ^2 is the noise power. Z_k^0 denotes the set of computing service UEs offloaded to AP k. Then, the completion delay for offloaded computing task is described as $t_i^0 = \frac{b_i^o}{r_{ik}^u} + \frac{w_i^o}{f_{ik}}$. Here, f_{ik} is the computation capability allocated to UE i at AP k. In summary, the delay of the computing task for UE i with regard to offloading decision is given as follows.

$$t_i^0 = \sum_{k\in\mathcal{K}} \lambda_{ik}^0 \left(\frac{b_i^0}{r_{ik}^u} + \frac{w_i^0}{f_{ik}} \right) + \left(1 - \sum_{k\in\mathcal{K}} \lambda_{ik}^0 \right) \frac{w_i^0}{F_l} \tag{2}$$

C. CACHING MODEL

In regard of content delivery service, a task of UE i is characterized by b_i^1, which denotes the size of the file requested in bits. The file can be either restored in remote core server, or be previously cached in any AP. This poses a problem for optimal caching allocation for UEs based on the limitation of storage resources. Let λ_{ik}^1 denotes the caching decision from UE i to AP k. Specifically, $\lambda_{ik}^1 = 1$ if desirable file of UE i is cached in AP k and 0 otherwise. Assume that a requesting file of any UE can only be cached in one AP. Therefore, $1 - \sum_{k\in\mathcal{K}} \lambda_{ik}^1 = 0$ means that the file for UE i is cached in core network.

The delay for files cached in remote server consists of wireline delay from core network to APs and the wireless downloading delay. On the other hand, if the file is cached in AP, the delay is only the downlink transmission delay from a certain AP to the UE. The downlink transmission rate from AP k to UE i is denoted as

$$r_{ik}^d = B^d \log_2 \left(1 + \frac{\frac{p_k^d}{|Z_k^1|} G_{ik}^d}{\sum\limits_{j\in\{Z_k^1 \setminus i\}} \frac{p_k^d}{|Z_k^1|} G_{jk}^d + \sigma^2} \right), \tag{3}$$

where p_k^d is the downlink transmission power of AP k, G_{ik}^d is the channel gain from AP k to UE i and Z_k^1 denotes the

set of content delivery service UEs associated to AP k, with $|Z_k^1|$ as the number of UEs in Z_k^1. Let r^b denotes the average transmission rate of wireline link. Then, the delay of the content delivery task of UE i in terms of caching allocation decision is expressed as follows.

$$t_i^1 = \sum_{k\in\mathcal{K}} \lambda_{ik}^1 \left(\frac{b_i^1}{r_{ik}^d}\right) + \left(1 - \sum_{k\in\mathcal{K}} \lambda_{ik}^1\right)\left(\frac{b_i^1}{r^b} + \frac{b_i^1}{r_{ik'}^d}\right) \quad (4)$$

Here, we assume that UEs get the files from the nearest AP k' if the requesting file is stored in core network and $r_{ik'}^d$ denotes the corresponding transmission speed.

D. PROBLEM FORMULATION

Both of the two services are deployed in the MEC network. Computing service UEs occupy both computation and storage resources in the associated APs. On the other hand, content delivery service UEs take the storage space if their requesting files are cached. Considering these two heterogeneous services together, there exists a challenge of resource allocation and association decision concerned with UEs. Let ω_i^c denotes the delay sensitive coefficient of UE i for service class c, similar with that in [27]. Larger value of ω_i^c illustrates the higher requirement of task delay. Thus, we formulate the weighted sum of delays of UEs shown as follows.

$$WT = \sum_{c=0}^{1} \sum_{i\in\mathcal{N}_c} \omega_i^c t_i^c. \quad (5)$$

WT can illustrate the performance of QoE in terms of service requirements and individual pays. Then, the network objective is to find the feasible resource allocation on computation and storage resources in MEC and the association assignment $\boldsymbol{\lambda} = \{\lambda_i^c, c = 0, 1, i \in \mathcal{N}_c\}$ to minimize the value of WT. This problem can be formulated as

$$\begin{aligned} \min_{\{\lambda_{ik}^c, f_{ik}\}} \quad & \sum_{c=0}^{1} \sum_{i\in\mathcal{N}_c} \omega_i^c t_i^c \\ s.t.\ C1: \quad & \sum_{k\in\mathcal{K}} \lambda_{ik}^c \le 1, \quad \forall i, c \\ C2: \quad & \sum_{i\in\mathcal{N}_0} \lambda_{ik}^0 f_{ik} \le F_m, \quad \forall k \\ C3: \quad & \sum_{c=0}^{1} \sum_{i\in\mathcal{N}_c} \lambda_{ik}^c b_i^c \le M, \quad \forall k. \end{aligned} \quad (6)$$

Constraint C1 ensures that one UE can associated to one AP at most. Constraints C2-C3 guarantee the resource limitations in MEC. The problem in (6) is MINLP, which is NP-hard to solve.

III. COALITION GAME APPROACH

In this section, coalition game is utilized to model the association assignment between APs and UEs. Then, a coalition game based algorithm for resource allocation and association decision is proposed.

A. COALITION GAME FORMULATION

In light of the formulated problem to minimize the weighted sum of delays, a coalition game model is introduced as UEs have incentives to form coalitions in order to decrease the value of WT. There are K APs with MEC and $|\mathcal{N}|$ UEs in the network, where UEs associated with the same AP share resources and form a coalition. Moreover, computing service UEs can execute locally while content files from content delivery service can be stored in core network, thus, we suppose that there are $(K + N_0 + 1)$ coalitions formed by UEs in the network. Let $F = \{F_1, ..., F_{K+N_0}, F_{K+N_0+1}\}$ denotes the collection of coalitions, where $F_x \cap F_y = \emptyset, \forall F_x, F_y \in F$, and $\cup_{x=1}^{K+N_0+1} F_x = \mathcal{N}$. The cardinality of F is the number of coalitions. In addition, for any $F_x \in F, x \in \{1, 2, ..., K\}$, the coalition is composed of the set of UEs associated with AP x. For any $F_x \in F, x \in \{K + 1, K + 2, ..., K + N_0\}$, the coalition is the local device UE x of computing service. For $F_{K+N_0+1} \in F$, the coalition is the set of UEs of content delivery service whose requesting files are cached in core network.

It can be observed that the larger the number of UEs associated with a certain AP of any service, the greater the transmission interference with less resources allocated individually. Processing locally for computing task or caching in remote server have no efforts to decrease WT as for larger completion delay. Therefore, there is no motivation for UEs to neither form a grand coalition nor complete tasks uncooperatively. This promotes the further investigation to explore the optimal coalition structure. In this paper, the optimal resource allocation and association assignment are modeled in a coalition formation game with transferable utility [24], [28], where UEs as game players, tend to form coalitions to improve the overall game utility. Then, we define the following coalition formation game with transferable utility.

Definition 1 (Coalition Game With Transferable Utility): A coalition game with transferable utility for resource allocation of heterogeneous services is defined by a pair $(\mathcal{N}, Q)$, where $\mathcal{N}$ is the set of players, and Q is the payoff function. For any coalition structure F, $Q(F)$ represents the network utility calculated with players cooperative structure F. Thus, the utility function of a partition F can be defined as

$$Q(F) = WT_{ind} - WT, \quad (7)$$

where WT_{ind} is the weighted sum of UEs' delays when all computing service UEs execute task processing locally and content delivery service UEs get the requesting files from the core network. Then, $WT_{ind} = \sum_{i\in\mathcal{N}_0} \omega_i^0 \frac{w_i^0}{F_l} + \sum_{i\in\mathcal{N}_1} \omega_i^1 \left(\frac{b_i^1}{r^b} + \frac{b_i^1}{r_{ik'}^d}\right)$. In addition, the utility of a certain coalition F_x is defined as follows.

$$Q(F_x) = WT_{ind,F_x} - \left(\sum_{i\in Z_x^0} \omega_i^0 t_i^0 + \sum_{i\in Z_x^1} \omega_i^1 t_i^1\right). \quad (8)$$

Here, WT_{ind,F_x} is the weighted sum of delays for the set of UEs in coalition k if they are not supported by MEC,

i.e., $WT_{ind,F_x} = \sum_{i\in Z_x^0} \omega_i^0 \frac{w_i^0}{F_l} + \sum_{i\in Z_x^1} \omega_i^1 \left(\frac{b_i^1}{r^b} + \frac{b_i^1}{r_{ik'}^d}\right)$. For any $F_x \in F$, the value of $Q(F_x)$ can be distributed to its members.

It is obvious that the greater payoff function is, the less WT defined in (5). Thus, the maximization of the utility $Q(F), F \in \mathbb{F}$ is equal to the optimization problem in (6). Further, we define a coalition formation game for resource allocation and UE's association of heterogeneous services according to the basics in [29].

Definition 2 (Coalition Formation Game for Resource Allocation of Services): The coalition formation game for resource allocation and UEs' association of heterogeneous services is defined by $(\mathcal{N}, Q, F)$, where F is the partition of all UEs and $F \in \mathbb{F}$. To be specific, the partition F is the collection of coalitions $S = \{S_1, ..., S_L\}$ where L is the number of the coalitions with $\cup_{l=1}^{L} S_l = \mathcal{N}$.

B. COALITION GAME BASED ALGORITHM

The essential ingredient of the coalition formation game is to design a well defined order to compare two partitions, and then set rules to enable players join or break their coalitions based on preference. Thus, we present the following preference relation for UEs.

Definition 3 (Preference Relation): The preference relation for any two partitions of the subset $A \subseteq \mathcal{N}$ is defined as $\succ$. $F \succ F'$ denotes that partition F is preferable than F' to the overall players of A. For any UE $i \in \mathcal{N}$, the preference relation is defined by $\succ_i$. $F_x \succ_i F_y$ denotes that partition F_x is preferable than F_y for UE i.

In this paper, UEs decide to join or leave a coalition in accordance with the utility of the partition. In summary, the preference for two partition F and F' of UEs can be defined as follows:

$$F \succ F' \Leftrightarrow Q(F) > Q(F'). \tag{9a}$$

$$F_x \succ_i F_y \Leftrightarrow Q(F_x') + Q(F_y') > Q(F_x) + Q(F_y),$$
$$F_x' = (F_x \cup \{i\}), \quad F_y' = (F_y \setminus \{i\}). \tag{9b}$$

It is observed that $F \succ F'$ is guaranteed by preference condition in (9b), where the new partition formed in (9b) is $F' = (F \setminus \{F_x, F_y\}) \cup F_x' \cup F_y'$. Furthermore, assume that the current partition of $\mathcal{N}$ is $F = \{F_1, ..., F_{K+N_0}, F_{K+N_0+1}\}$, we define three operations for UEs of different services to join or leave the coalitions, shown as follows:

- *Merge Operation:* For computing service UE i, the merge operation starts only if its coalition is F_{K+i}, while UE i is able to join the new coalition $F_k, k \leq K$. The preference is only influenced by the original coalition and new coalition. Thus, the merge operation is operated if $F_k \succ_i F_{K+i}, k \leq K$. Similarly, for content delivery service UE i, the merge operation from its coalition F_{K+N_0+1} to new coalition $F_k, k \leq K$ is operated if $F_k \succ_i F_{K+N_0+1}, k \leq K$.
- *Split Operation:* For computing service UE i, the split operation starts only when comparing its current coalition $F_k, k \leq K$ and new coalition F_{K+i}. The preference is only influenced by the original coalition and destination coalition. Thus, the split operation is operated if $F_{K+i} \succ_i F_k, k \leq K$. Similarly, for content delivery service UE i, the split operation from its current coalition $F_k, k \leq K$ to new coalition F_{K+N_0+1} is operated if $F_{K+N_0+1} \succ_i F_k, k \leq K$.
- *Exchange Operation:* Assume that any two UEs $i, j \in \mathcal{N}$ with their corresponding coalition $F_x, F_y, x, y \leq K$. A new partition of $\mathcal{N}$ is denoted as $F' = (F \setminus \{F_x, F_y\}) \cup \{F_x \setminus \{i\} \cup \{j\}, F_y \setminus \{j\} \cup \{i\}\}$. The exchange operation is operated for UE i and j if $F \succ F'$.

Algorithm 1 Coalition Game Based Algorithm

Initialization: Develop a random partition F_{ini}.
Let $F_{cur} = F_{ini}$. Set $n = 2$, $T_n = T_0$.
while not converges to Nash-stable **do**
  Uniformly randomly choose user i and its coalition F_x.
  //Merge and Split Operations
  if $i \in \mathcal{N}_0$ and $F_x = F_{K+i}$ **then**
    Uniformly randomly choose a new coalition $F_y, y \leq K$.
  end if
  if $i \in \mathcal{N}_0$ and $F_x \neq F_{K+i}$ **then**
    Let $F_y = F_{K+i}$.
  end if
  if $i \in \mathcal{N}_1$ and $F_x = F_{K+N_0+1}$ **then**
    Uniformly randomly choose a new coalition $F_y, y \leq K$.
  end if
  if $i \in \mathcal{N}_1$ and $F_x \neq F_{K+N_0+1}$ **then**
    Let $F_y = F_{K+N_0+1}$.
  end if
  Let $F' = (F \setminus \{F_x, F_y\}) \cup \{F_x \setminus \{i\}, F_y \cup \{i\}\}$.
  if $F' \succ F$ **then**
    Update $F_{cur} = F'$.
  else
    if $F \succ F'$ and $\rho_{F,F'}(T_n) > rand$ **then**
      Update $F_{cur} = F'$.
    end if
  end if
  //Exchange Operation
  Uniformly randomly choose two users i, j and their corresponding coalitions $F_x, F_y, x, y \leq K$.
  Let $F' = (F \setminus \{F_x, F_y\}) \cup \{F_x \setminus \{i\} \cup \{j\}, F_y \setminus \{j\} \cup \{i\}\}$.
  if $F' \succ F$ **then**
    Update $F_{cur} = F'$.
  else
    if $F \succ F'$ and $\rho_{F,F'}(T_n) > rand$ **then**
      Update $F_{cur} = F'$.
    end if
  end if
  Let $n = n + 1$; $T_n = \frac{T_0}{\log(n-1)}$.
end while

Operations are only feasibly executed when constraints C1-C3 are satisfied. Moreover, there is a chance for carrying out the operations if the new partition is not preferred in each operation. Based on the method of simulated annealing, an acceptance probability for changing to new partition is designed as

$$\rho_{F,F'}(T_n) = e^{\frac{Q(F')-Q(F)}{T_n}}, \tag{10}$$

where $T_n = \frac{T_0}{\log(n-1)}$. T_0 is the initial temperature in simulated annealing, and n is the current number of iterations. The acceptance probability $\rho_{F,F'}(T_n)$ is designed for avoidance of the premature for convergence, which may turn out to be a local optimum for the optimization problem.

In each iteration, resource allocation of computation resources in APs are investigated after every operation. With deterministic association assignment $\boldsymbol{\lambda}$ after the execution of any operation, the optimization problem can be solved independently among APs, which is shown as follows.

$$\begin{aligned} &\min_{\{f_{ik}\}} \sum_{i\in F_k} \omega_i^0 \left(\frac{b_i^0}{r_{ik}^u} + \frac{w_i^0}{f_{ik}}\right) \\ &s.t. \sum_{i\in F_k} f_{ik} \le F_m \quad \forall k. \end{aligned} \tag{11}$$

The optimization of resource allocation is convex to $\boldsymbol{f_k} = \{f_{ik}, i \in F_k\}, k \le K$. Therefore, it is feasibly solved by KKT conditions. The Lagrangian function can be presented as

$$L(\boldsymbol{f_k}, \lambda) = \sum_{i\in F_k} \omega_i^0 \left(\frac{b_i^0}{r_{ik}^u} + \frac{w_i^0}{f_{ik}}\right) + \lambda \left(\sum_{i\in F_k} f_{ik} - F_m\right), \tag{12}$$

where λ is the variable of nonnegative Lagrangian multiplier. Then, the optimal λ and $\boldsymbol{f_k}$ must satisfy the following equalities:

$$\frac{dL(\boldsymbol{f_k}, \lambda)}{df_{ik}} = -\frac{\omega_i^0 w_i^0}{{f_{ik}}^2} + \lambda = 0 \tag{13}$$

$$\frac{dL(\boldsymbol{f_k}, \lambda)}{d\lambda} = \sum_{i\in F_k} f_{ik} - F_m = 0. \tag{14}$$

Thus, the optimal solution of computation resource is shown as

$$\lambda^* = \frac{\left(\sum_{i\in F_k} \sqrt{\omega_i^0 w_i^0}\right)^2}{{F_m}^2} \tag{15}$$

$$f_{ik}{}^* = \frac{F_m \sqrt{\omega_i^0 w_i^0}}{\sum_{i\in F_k} \sqrt{\omega_i^0 w_i^0}}, \quad \forall k, i. \tag{16}$$

The coalition game based algorithm for resource allocation and association assignment with heterogeneous services is summarized in Algorithm 1. In each iteration, the system randomly chooses an UE i for merge or split operation. The information of its current coalition F_x and a new coalition F_y is got as shown in Line 6-17, which are then compared with each other according to preference relation. If new partition after the operation are preferable, the operation is carried out. Otherwise, the operation is executed within an acceptance probability related to the number of iterations completed, shown in Line 19-25. Then, the system uniformly choose two UEs i, j with their corresponding coalitions F_x and F_y for exchange operation as shown in Line 27. Coalitions F_x and F_y exchange their member i and j to form a new partition F' in Line 28. If the new partition after exchanging is preferable, the exchange operation is carried out. Otherwise, it still works out within an acceptance probability, as shown in Line 29-35. The algorithm ends if converges to a Nash-stable solution or reaches the maximal number of iterations.

IV. THEORETIC ANALYSIS

In this section, the stability of the proposed algorithm is analyzed. Then, the optimality is presented based on Markov chain theory. Finally, the convergence of the proposed algorithm is given.

A. STABILITY

Definition 4 (Nash-Stable Structure): A coalition structure is Nash-stable if $\forall i \in \mathcal{N}$, no further operations are carried out.

Theorem 1: The final coalition structure F_{fin} in Algorithm 1 is Nash-stable.

Proof: In every iteration of Algorithm 1, the partition is either reformed to a new structure or stays the same. Since the number of coalition is $K + N_0 + 1$, the maximal number of coalition structures is finite. Thus, after a certain number of operation with merging, splitting and exchanging, the change of partition is terminated and converged to the a final structure. If the final partition F_{fin} of Algorithm 1 is not Nash-stable, there exists $i \in \mathcal{N}$ with its coalition F_x, and $F_y \in F_{fin}, F_y \ne F_x$ such that F_y is preferred to i than F_x, or there exist $i, j \in \mathcal{N}$ with their corresponding coalitions F_x and F_y such that $Q(F') > Q(F_{fin})$ where $F' = \{F_{fin} \backslash \{F_x, F_y\} \cup \{F_x \backslash \{i\} \cup \{j\}, F_y \backslash \{j\} \cup \{i\}\}$. According to the definition of operations, one of merging, splitting and exchanging operations is made definitely. This is contrary to the definition that F_{fin} is the final partition. Therefore, the final partition F_{fin} obtained by Algorithm 1 is a Nash-stable coalition structure. □

B. OPTIMALITY

Theorem 2: The solution obtained by Algorithm 1 converges to optimality with increasing number of operations.

Proof: The process of evolvement of the partition of UEs can be seen as a Markov chain in terms of parameter T_n. Then, we first prove that the Markov chain $\{F(T_n)\}$ is ergodic. Denote the current partition and new partition as F and F', respectively. Referring to Algorithm 1, when F' is preferred rather than F, that is, $Q(F') > Q(F)$, the transition of partition is accepted with probability 1. Whereas, the partition changes to the worse one with a probability $\rho_{F,F'}(T_n)$. Obviously, the probability of transiting to F' that is worse than F is

reduced as $|Q(F') - Q(F)|$ increases. Thus, the probability of transition of the partition in the n_{th} phase is expressed as follows.

$$\rho_{F,F'}(T_n) = \begin{cases} 1 & Q(F) < Q(F') \\ e^{\frac{Q(F') - Q(F)}{T_n}} & Q(F) \geq Q(F'). \end{cases} \tag{17}$$

Then, we can obtain that

$$\lim_{T_n \to 0} \rho_{F,F'}(T_n) = \begin{cases} 1 & Q(F) < Q(F') \\ 1 & Q(F) = Q(F') \\ 0 & Q(F) > Q(F'). \end{cases} \tag{18}$$

This implies that the proposed algorithm is less permissive as T_n reaches 0. We define

$$\underline{\rho_{F,F'}}(T_n) = \inf_{F \in \mathbb{F}, F' \in N(F)} \rho_{F,F'}(T_n), \tag{19}$$

where $N(F)$ is the set of neighbors of F. According to (17), $\underline{\rho_{F,F'}}(T_n)$ can be expressed as

$$\begin{aligned} \underline{\rho_{F,F'}}(T_n) &= \inf_{\substack{F,F' \in \mathbb{F} \\ F' \neq F}} \rho_{F,F'}(T_n) \\ &= \inf_{\substack{F,F' \in \mathbb{F} \\ Q(F) > Q(F')}} e^{\frac{Q(F') - Q(F)}{T_n}} \geq e^{-\frac{\Delta}{T_n}} \end{aligned} \tag{20}$$

where $\Delta = \sup\{Q(F) - Q(F'), F' \in N(F)\}$. Since Δ is a constant in the process, we can set $T_0 \leq G\Delta$, where G is a relatively large constant. Moreover, as $T_n = \frac{T_0}{\log(n-1)}$, the cooling schedule $\{T_n\}_{n>0}$ in simulated annealing satisfies

$$T_n \leq \frac{T_0}{\log(n)}. \tag{21}$$

Thus, we can have

$$\begin{aligned} \sum_{n=1}^{\infty} \left(\underline{\rho_{F,F'}}(T_{nG})\right)^G &\geq \sum_{n=1}^{\infty} e^{\frac{-G\Delta}{T_n}} \\ &\geq \sum_{n=1}^{\infty} e^{\frac{G\Delta}{T_0} \log \frac{1}{nG}} \\ &\geq \sum_{n=1}^{\infty} \frac{1}{nG} = \infty. \end{aligned} \tag{22}$$

Based on [30, Ch. 7, Th. 8.1], $\{F(T_n)\}$ is weakly ergodic. Furthermore, in view of Theorem 8.2 of Chapter 6 in [30], $\{F(T_n)\}$ is strongly ergodic. Also, $\{F(T_n)\}$ is irreducible, thus, the stationary distribution exists and is unique, which is equal to the limiting probability as n sufficiently large of Algorithm 1. Denote the stationary distribution of $\{F(T_n)\}$ as $\pi(F)$, thus, the stationary distribution is shown as

$$\pi(F) = \frac{e^{Q(F)/T_n}}{\sum_{F' \in \mathbb{F}} e^{Q(F')/T_n}} \tag{23}$$

In the next step, we verify that the limiting probability vector, which is the same as the stationary distribution, pulls all its mass on the set of global maxima of the utility function $Q(\cdot)$, which equally reaches the set of minima of the original problem. Define the set of coalition structures for global minima of $Q(\cdot)$ as H, which is expressed as follows,

$$H = \{X \in \mathbb{F}; Q(X) \geq Q(Y), \forall Y \in \mathbb{F}\}. \tag{24}$$

Define the maxima of transferable utility of coalition game as $m = \max_{X \in \mathbb{F}} Q(X)$. By dividing the numerator and denominator of $\pi(F)$ in (23) by $e^{-\frac{m}{T_n}}$, then, we can rewrite the stationary distribution as

$$\begin{aligned} \pi(F) &= \frac{e^{-\frac{m-Q(F)}{T_n}}}{\sum_{F' \in \mathbb{F}} -\frac{m-Q(F')}{T_n}} \\ &= \frac{e^{-\frac{m-Q(F)}{T_n}}}{|H| + \sum_{F' \in \mathbb{F}, F' \notin H} -\frac{m-Q(F')}{T_n}}. \end{aligned} \tag{25}$$

As the number of operations n increases, $T_n \to 0$. Thus, we get

$$\lim_{T_n \to 0} e^{-\frac{m-Q(F')}{T_n}} = \begin{cases} 1 & F' \in H \\ 0 & F' \notin H. \end{cases} \tag{26}$$

Then, the stationary distribution when $T_n \to 0$ is equal to

$$\lim_{T_n \to 0} \pi(F) = \begin{cases} \frac{1}{|H|} & F \in H \\ 0 & F \notin H. \end{cases} \tag{27}$$

This shows that the steady state converges to the global maxima of the coalition game with probability 1. As the maxima of coalition game is equal to the minima of the original problem in (6), Algorithm 1 is verified to converge to the set of global optimum, which minimize the weighted sum of delays of UEs in the network. □

C. CONVERGENCE

In each iteration of Algorithm 1, one UE is chose such that the utilities of the current associated coalition and another chose coalition are calculated. Then, a merging or splitting operation is performed with the probability in (17). Similarly, exchanging decision is made based on the coalitions of the two randomly chose users. Consequently, there are almost two operations in each iteration. The fast convergence performance of the proposed algorithm is shown in Figure 10 and section V.F, which show the efficiency of the proposed algorithm.

V. SIMULATION RESULTS

In this section, we evaluate the simulation results of the proposed Coalition Game Based Algorithm, referred as *CGBA*, for resource allocation and association assignment of heterogeneous services. We consider a network with area size 200×200. There exist $|\mathcal{K}| = 10$ APs and $|\mathcal{N}| = 50$ UEs, which are uniformly randomly distributed in this area. The group of UEs for two services are divided by the parameter user ratio $\alpha = 0.5, 0 \leq \alpha \leq 1$, $|\mathcal{N}_0| = \lfloor \alpha |\mathcal{N}| \rfloor$, and $|\mathcal{N}_1| = \lceil (1-\alpha)|\mathcal{N}| \rceil$. Here, $\alpha = 0.5$. The size of the input b_i^0 has uniform distribution within the range [10, 100] Mbits, and the computation requirement $w_i^0 = 2000 \times b_i^0$ for computing

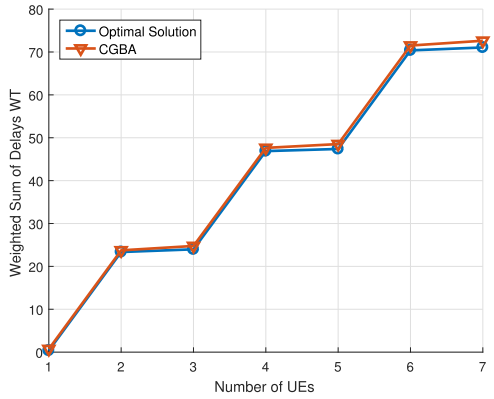

FIGURE 2. Performance of the weighted sum of delays with increasing number of UEs in terms of CGBA and the optimal solution.

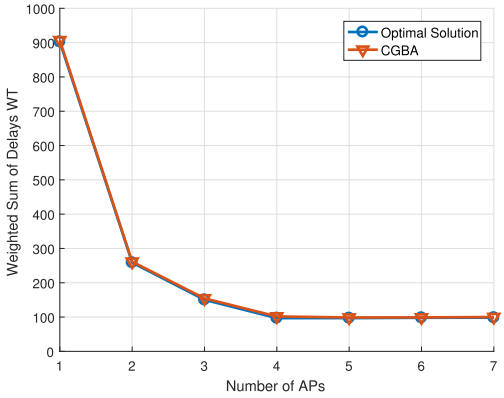

FIGURE 3. Performance of the weighted sum of delays with increasing number of APs in terms of CGBA and the optimal solution.

service UEs. The size of the file b_i^1 is uniformly distributed within the range [100, 500] Mbits. The transmission power of UEs and APs are $p_i^u = 20$ dBm and $p_k^d = 30$ dBm, respectively. The noise power is $\sigma^2 = 2.0 \times 10^{-11}$ W. The computation capability and storage capability in each AP are $F_m = 5 \times 10^9$ CPU cycles/s and $M = 2$ GB, respectively. The local computation capability is $F_l = 0.5 \times 10^9$ CPU cycles/s. We set $\omega_i^c = 1, \forall c, i$.

A. PERFORMANCE COMPARISON WITH OPTIMAL SOLUTION

In this section, we present that our proposed CGBA gets the solution of the weighted sum of delays close to the optimum. The optimal solution is generated by exhaustive search for optimal solution of resource allocation and association assignment. The complexity of optimality search is $O\left((K+1)^{|\mathcal{N}|}\right)$, which exponentially increases with the number of APs and UEs. We firstly set $K = 5$ and vary $|\mathcal{N}|$ from 1 to 7. The performance comparison of the weighted sum of delays between CGBA and the optimal solution is shown in Fig.2. Then, we evaluate the performance with $|\mathcal{N}| = 8$ and the number of APs ranging from 1 to 7, as shown in Fig.3. From these two figures, it is observed that the performance of CGBA is almost the same to the optimal solution. Further, we calculate the average deviation from CGBA to the optimum in the cases of these two figures. The average deviation is defined as $D = \frac{1}{7}\sum_{x=1}^{7} \frac{WT^{CGBA}(x) - WT^{OP}(x)}{WT^{OP}(x)}$, where $WT^{CGBA}(x)$ and $WT^{OP}(x)$ denotes the weighted sum of delays of all UEs by CGBA and exhaustive search with different parameter x, respectively. Here, x denotes either the number of UEs in Figure 2 or the number of APs in Figure 3. In consequence, the average deviation $D = 15.9\%$ with different number of UEs and $D = 2.0\%$ with different number of APs.

B. PERFORMANCE COMPARISON WITH OTHER SCHEMES

In this section, the performance of the proposed CGBA is evaluated and compared with the following schemes:

1) PRiority ASsignment (PRAS), which is a modified version of the scheme in our former work in [31]. For all UEs, their priorities are the same as the task delay without MEC. UEs are able to choose their best APs for assignment based on the priorities, for example, the best choice for UE i to associate with is $k^* = \text{argmin}_{k \in \mathcal{K}} \omega_i^c t_i^c, \forall c$. Resource allocation is optimized each time an UE is assigned to a certain AP.
2) NEarest ASsignment (NEAS), where UEs associate with their corresponding nearest APs in a random order. If the AP is overloaded without extra resources, following UEs will either compute locally or get files from the core network.
3) WithOut MEC (WOMEC), where computing service UEs execute locally, and content delivery UEs get the desirable files from the core network.

Fig.4 presents the comparison of CGBA to the other schemes mentioned above in terms of the weighted sum of delays of UEs. It can be observed that CGBA, PRAS and NEAS all make efforts to reduce the weighted sum of delays WT with increasing amount of computing resources in MEC. Our proposed CGBA outperforms compared with the other scheme with much lower value of WT and continuous decreasing trend. Moreover, Fig.5 investigates the results of the weighted sum of delays with increasing number of UEs for CGBA and the other schemes. The proposed CGBA achieves the lowest value of WT under different number of UEs in the network. In addition, WT of CGBA increases slowly compared to those of the other schemes, which means that with larger number of UEs, our proposed algorithm makes more impacts on task delay reduction compared to PRAS, NEAS and WOMEC. Specifically, CGBA reduces the weighted sum of delays with average 27.8%, 82.1% and 85.5% to those of PRAS, NEAS, WOMEC, respectively.

C. FAIRNESS COMPARISON OF CGBA WITH OTHER SCHEMES

In this section, the fairness of resource allocation for UEs in the network is further discussed. The performance of fairness is an ignorable issue to evaluate the optimization of resource allocation [32], since all UE expect better QoE performance.

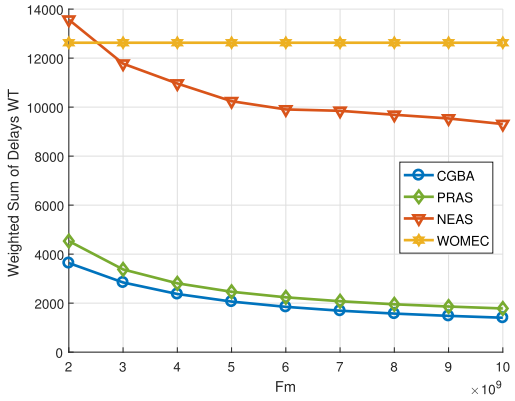

FIGURE 4. Performance on the weighted sum of delays of UEs of CGBA in the comparison with the other schemes with different amount of resources in MEC.

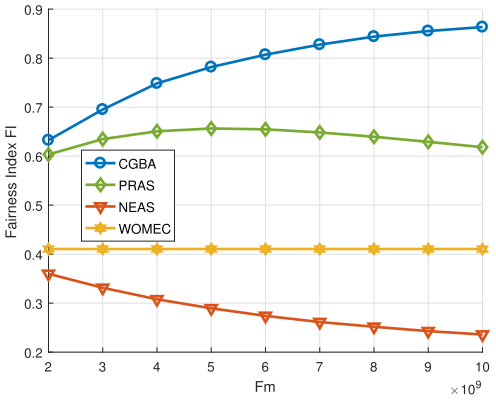

FIGURE 6. Fairness performance of CGBA versus different amount of computation resources in comparison to the other schemes.

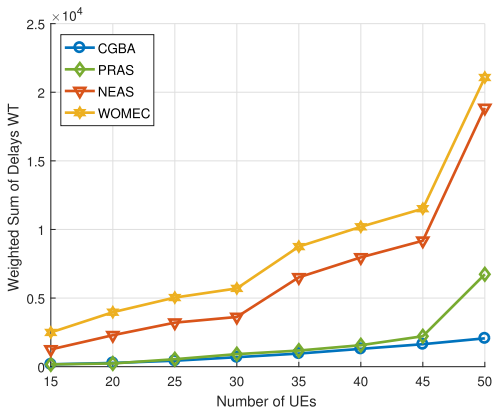

FIGURE 5. Performance on the weighted sum of delays of UEs of CGBA in the comparison with the other schemes with different number of UEs.

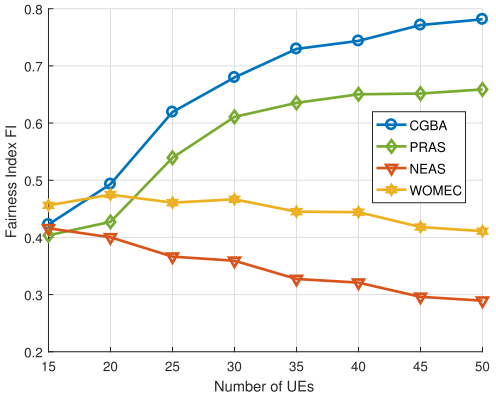

FIGURE 7. Fairness performance of CGBA versus different number of UEs in comparison to the other schemes.

As we optimize the delay-aware performance of UEs, the task delay is a main concern for UEs. Therefore, we evaluate the fairness performance in terms of task delay. Similar with in [33] and [34], we adopt the fairness index denoted as

$$FI = \frac{\left(\sum_{c=0}^{1}\sum_{i\in\mathcal{N}_c} t_i^c\right)^2}{|\mathcal{N}|\sum_c\sum_i \left(t_i^c\right)^2}. \tag{28}$$

FI locates in the range (0, 1], and higher value of F indicates better performance of fairness.

In Fig.6, we evaluate the results of fairness index FI with different amount of computation resources in terms of CGBA and the other schemes for comparison. It is observed that as the amount of resources increases, NEAS has worse performance of fairness, and PRAS cannot always make impacts on the improvement of fairness. It is because that NEAS neglects the feasible resources at the accessible APs beyond the nearest one, and for PRAS, the UEs with higher priorities are given more resources than those with lower priorities. On the other hand, CGBA gets higher fairness index as F_m increases. The increment of F_m increases the computing capability of all accessible APs besides the nearest one. Since CGBA has a flexible and optimal user association scheme, users are allocated more sufficient resources as F_m increases, which increases the users' fairness index FI.

Fig.7 depicts the fairness index with increasing number of UEs. The larger number of UEs, the higher load in the network. It is shown that NEAS worsens the fairness performance when changing the number of UEs due to its fixed association scheme. CGBA and PRAS are shown to increase the fairness index with increasing number of UEs. Moreover, our proposed CGBA gets the largest fairness index compared with the other. Both of the figures indicates that our proposed algorithm achieves good delay-aware performance with the fair resource allocation mechanism in accordance with the results in Fig.4 and Fig.5.

D. RESOURCE ALLOCATION FOR HETEROGENEOUS SERVICES

As mentioned before, computing service and content delivery service compete on storage resources in MEC, and this pose the problem for storage resource allocation in the network. In this section, we investigate the resource allocated to heterogeneous services concerned with delay sensitive coefficients and user distribution. Fig.8 presents the resource allocation ratio allocated to computing service versus the difference of

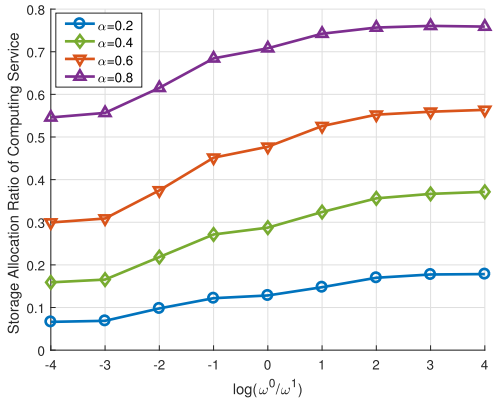

FIGURE 8. Storage resource allocation ratio of computing service in terms of different user ratio α.

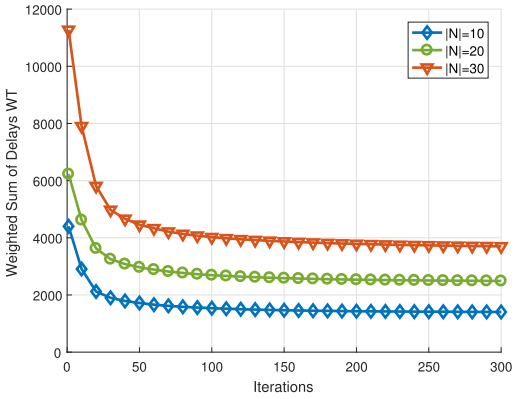

FIGURE 10. The performance of CGBA with the number of iterations. CGBA can converge in a small number of iterations.

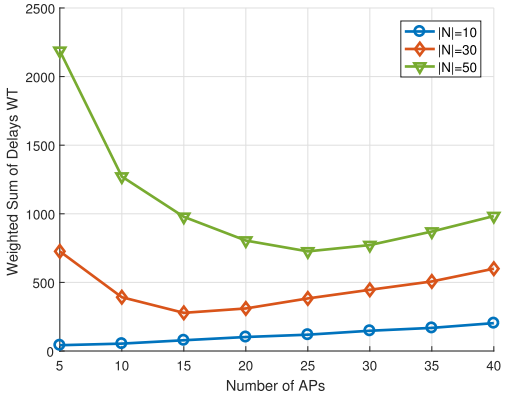

FIGURE 9. The performance on the weighted sum of delays in terms of different $|\mathcal{N}|$ versus different $|\mathcal{K}|$.

delay sensitive coefficients of two services. Here, we assume that $\omega^0 = \omega_i^0, \forall i \in \mathcal{N}_0$, $\omega^1 = \omega_i^1, \forall i \in \mathcal{N}_0$. The vertical axis is calculated as $\frac{M^0}{M^0+M^1}$, where M^0 and M^1 denote the sum of storage resources allocated to computing service and content delivery service, respectively. It is shown that as the number of computing service UEs increases, the storage allocated to this service increases even with a lower delay sensitive coefficient. This is reasonable according to the pursuit of WT minimization.

E. CGBA's PERFORMANCE UNDER DIFFERENT $|\mathcal{N}|$ AND $|\mathcal{K}|$

In this section, the performance on the weighted sum of delay in terms of different number of UEs $|\mathcal{N}|$ corresponding to varying number of APs $|\mathcal{K}|$ is analysed. In this experiment, we assume that the total amount of computation capability, storage resources and bandwidth resources are fixed. With varying number of APs, the resources are equally divided to each AP. For example, assume that the total amount of computation, storage and bandwidth resources are 60 GCPU cycles/s, 25 GB and 20 MHz, respectively. The performance with these parameters are shown in Fig.9. It is observed that if the number of UEs is 10, increasing the number of APs from 5 to 40 cannot reduce the weighted sum of delay. However, when increasing $|\mathcal{N}|$, there is a best choice on $|\mathcal{K}|$ for the minima of WT, for example, 15 if $|\mathcal{N}| = 30$ and 25 if $|\mathcal{N}| = 30$. This demonstrates that the distribution of APs with regard to the number of UEs is a significant issue for network performance, which we may study in the future work.

F. CONVERGENCE RATE EVALUATION

In this section, the convergence rate of the proposed CGBA is evaluated, as shown in Fig.10. It is shown that with the number of UEs increasing from 10 to 30, our proposed algorithm can still converge within at most 200 iterations. As a reference, the complexity of exhaustive search is $O(11^{30})$ if $\mathcal{N} = 30$, which is much larger than the number of iterations of the proposed CGBA for convergence. This further demonstrates the theoretical analysis in Section IV.

VI. CONCLUSION

In this paper, we investigate the issue of resource allocation when heterogeneous services work collaboratively in the MEC network. Then, we formulate an optimization problem of resource allocation to minimize the weighted sum of users' delays in accordance with association assignment. As the optimization problem is NP hard, a coalition game based algorithm is proposed to efficiently optimize the assignment vector and further resource management. Both theoretical analysis and numerical results show that the proposed algorithm converges to be Nash-stable fast. Further, simulation results also present the superior performance of QoE and fairness compared with other schemes. In the future, we may further take the distribution of APs into the optimization problem under the conditions of UEs, or exploit resource management in a multitude MEC architecture.

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] Z. Tan, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Virtual resource allocation for heterogeneous services in full duplex-enabled SCNs with mobile edge computing and caching," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1794–1808, Feb. 2018.

[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] ABI Research Lab. *Augmented and Virtual Reality: The First Wave of 5G Killer Apps*. Accessed: Feb. 2017. [Online]. Available: https://www.abiresearch.com/whitepapers/augmented-and-virtual-reality-first-wave-5g-killer/
[5] European Telecommunications Standards Institute. *MEC in 5G Networks*. Accessed: Jun. 2018. [Online]. Available: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf
[6] 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, 2017.
[7] 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.
[8] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397–1411, Mar. 2017.
[9] O. Muñoz, A. Pascual-Iserte, and J. Vidal, "Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4738–4755, Oct. 2015.
[10] T. Q. Dinh, J. Tang, Q. D. 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.
[11] J. Ren, G. Yu, Y. Cai, and Y. He, "Latency optimization for resource allocation in mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5506–5519, Aug. 2018.
[12] T. Yang, H. Zhang, H. Ji, and X. Li, "Computation collaboration in ultra dense network integrated with mobile edge computing," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Montreal, QC, Canada, Oct. 2017, pp. 1–5.
[13] T. Zhao, S. Zhou, X. Guo, Y. Zhao, and Z. Niu, "A cooperative scheduling scheme of local cloud and Internet cloud for delay-aware mobile cloud computing," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
[14] X. Ma, S. Zhang, P. Yang, N. Zhang, C. Lin, and X. Shen, "Cost-efficient resource provisioning in cloud assisted mobile edge computing," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017, pp. 1–6.
[15] C. Liang, Y. He, F. R. Yu, and N. Zhao, "Enhancing QoE-aware wireless edge caching with software-defined wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6912–6925, Oct. 2017.
[16] C. Li, L. Toni, J. Zou, H. Xiong, and P. Frossard, "QoE-driven mobile edge caching placement for adaptive video streaming," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 965–984, Apr. 2018.
[17] T. Hou, G. Feng, S. Qin, and W. Jiang, "Proactive content caching by exploiting transfer learning for mobile edge computing," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017, pp. 1–6.
[18] X. Zhang and Q. Zhu, "Collaborative hierarchical caching over 5G edge computing mobile wireless networks," in *Proc. ICC*, Kansas City, MO, USA, May 2018, pp. 1–6.
[19] C. Yi, S. Huang, and J. Cai, "An incentive mechanism integrating joint power, channel and link management for social-aware D2D content sharing and proactive caching," *IEEE Trans. Mobile Comput.*, vol. 17, no. 4, pp. 789–802, Apr. 2018.
[20] X. Peng, J. Zhang, S. H. Song, and K. B. Letaief, "Cache size allocation in backhaul limited wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[21] H. Wang, R. Li, L. Fan, and H. Zhang, "Joint computation offloading and data caching with delay optimization in mobile-edge computing systems," in *Proc. Int. Conf. Wireless Commun. Signal Process.*, Nanjing, China, Oct. 2017, pp. 1–6.
[22] F. Wang, Y. Li, Z. Wang, and Z. Yang, "Social-community-aware resource allocation for D2D communications underlaying cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3628–3640, May 2016.
[23] S.-C. Zhan and D. Niyato, "A coalition formation game for remote radio head cooperation in cloud radio access network," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1723–1738, Feb. 2017.
[24] Y. Chen, B. Ai, Y. Niu, K. Guan, and Z. Han, "Resource allocation for device-to-device communications underlaying heterogeneous cellular networks using coalitional games," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4163–4176, Jun. 2018.
[25] S. Bi and Y. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4177–4190, Jun. 2018.
[26] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, Aug. 2017.
[27] M. R. Mardani, S. Mohebi, B. Maham, and M. Bennis, "Delay-sensitive resource allocation for relay-aided M2M communication over LTE-advanced networks," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Heraklion, Greece, Jul. 2017, pp. 1033–1038.
[28] D. Wu, Y. Cai, L. Zhou, and J. Wang, "A cooperative communication scheme based on coalition formation game in clustered wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1190–1200, Mar. 2012.
[29] Z. Han, D. Niyato, W. Saad, and A. Hjorungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[30] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. New York, NY, USA: Springer-Verlag, 1999.
[31] J. Zhou, C. Sun, X. Zhang, and W. Wang, "Delay-aware resource management for heterogeneous service collaboration in mobile edge networks," in *Proc. IEEE ICC*, Kansas City, MO, USA, May 2018, pp. 1–6.
[32] 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.
[33] Y. Yang and T.-S. P. Yum, "Multicode multirate compact assignment of OVSF codes for QoS differentiated terminals," *IEEE Trans. Veh. Technol.*, vol. 54, no. 6, pp. 2114–2124, Nov. 2005.
[34] S. Zhao, Y. Yang, Z. Shao, X. Yang, H. Qian, and C.-X. Wang, "FEMOS: Fog-enabled multitier operations scheduling in dynamic wireless networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1169–1183, Apr. 2018.

JIZHE ZHOU received the B.S. degree in electronic and information engineering from Beihang University, China, in 2015, and the M.S. degree in computer science from Columbia University, USA, in 2016. She is currently pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communications, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include 5G network technology, mobile edge computing, and caching technology.

XING ZHANG is currently a Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. His current research interests include satellite networks, mobile edge computing, wireless big data, and the Internet of Things.

WENBO WANG received the B.S., M.S., and Ph.D. degrees from the Beijing University of Posts and Telecommunications, in 1986, 1989, and 1992, respectively, where he is currently the Assistant Director of the Key Laboratory of Universal Wireless Communication, Ministry of Education. He has published more than 200 journal and international conference papers and six books. His current research interests include radio transmission technology, wireless network theory, and software radio technology.