Double-Matching Resource Allocation Strategy in Fog Computing Networks Based on Cost Efficiency

Boqi Jia, Honglin Hu, Yu Zeng, Tianheng Xu, and Yang Yang

Abstract: **Fog computing is an advanced technique to decrease latency and network congestion, and provide economical gains for Internet of Things (IoT) networks. In this paper, we investigate the computing resource allocation problem in three-layer fog computing networks. We first formulated the resource allocation problem as a double two-sided matching optimization problem. Then, we propose a double-matching strategy for the resource allocation problem in fog computing networks based on cost efficiency, which is derived by analysing the utility and cost in fog computing networks. The proposed double-matching strategy is an extension of the deferred acceptance algorithm from two-side matching to three-side matching. Numerical results show that high cost efficiency performance can be achieved by adopting the proposed strategy. Furthermore, by using the proposed strategy, the three participants in the fog computing networks could achieve stable results that each participant cannot change its paired partner unilaterally for more cost efficiency.**

Index Terms: **Cost efficiency, fog computing networks, matching, resource allocation.**

I. INTRODUCTION

THE Internet of Things (IoT) is one of the hottest megatrends in technology, and the number of connected units has increased including 26 billion devices installed by 2020 at a unprecedented speed [1]. IoT deployments will generate large quantities of data that need to be processed and analyzed in real time. Processing and analyzing the big IoT data in real time will increase as a proportion of workloads of data centers, requiring a large amount of computing resources and leaving providers facing new capacity and analytic challenges [2].

In order to deal with the challenges and meet the demand of the computing services, a large number of large-scale data centers or clouds has been deployed, and cloud computing has been proposed to provide flexible and efficient services to the users. In cloud computing, the cloud data centers is able to organize a shared pool of configurable computing resources (such as servers, storage, networks, applications, and services), which can be easily accessed by users on demands [3]. However, clouds are typically built in remote areas far from the users, which results in high transmission cost, transmission congestion, and service latency. Furthermore, it is intolerable for the applications that require real-time interaction in IoT scenarios.

Accordingly, it is beneficial and necessary to pull the cloud closer to the users. In IoT, fog computing [4]–[6], coined by Cisco, is proposed as a promising solution. Fog computing only pushes relevant data to the data centers and allows computing, decision-making and action-taking to happen via multiple low-power computing devices. These devices, called fog nodes (FNs), at the edge of the networks are deployed to offload the data computing services from the cloud. Therefore fog computing can provide low-latency, fast-response, and location-awareness service.

Fog computing networks consist of a large number of FNs deployed by different cloud data centers at different locations to provide various data services and applications to the users. Furthermore, network virtualization is applied in fog computing, where FNs are invisible for the users, the users can only purchase the computing resources from cloud data centers to process their excessive workload. When receiving requests from all users, each cloud data center is able to collect the computing resources from all the FNs. Note that each cloud data center can allocate different amount of computing resources from different FNs to different users, and hence the computing resource can be utilized by users efficiently. However, how to allocate the limited computing resources of all the FNs to the users is still an open problem.

In this paper, we focus on the computing resource allocation problem in fog computing networks and the main objective of this paper is to develop resource allocation strategy to maximize the cost efficiency. We organize our paper as follows. Related work is addressed in Section II. In Section III, we describe the fog computing network model. In Section IV, we derive the cost efficiency. In Section V, we present the resource allocation strategy, and in Section VI we show numerical results. Finally, Section VII concludes the paper.

II. RELATED WORK

Fog computing emerged in 2012 is a relatively recent research trend [7], [8]. In fog computing networks, efficiently allocating the cloud/fog computational resources to various users with various requirements has been studied in many literatures.

At the beginning, the service quality has been a key metric to measure various providers, specially in the case of mobile operators, since plenty of personal or business applications are migrated to cloud data centers in fog computing networks and fog

Manuscript received December 14, 2017.

This research was supported by the National Natural Science Foundation of China (Nos. 61671437, 61631013), the Program of Shanghai Academic Research Leader (No. 18XD1404100), and the ESEC project of Tekes.

B. Jia and Y. Zeng are with Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, China, email: {boqi.jia, yu.zeng}@wico.sh.

H. Hu and T. Xu are with Shanghai Advanced Research Institute, Chinese Academy of Sciences, China, email: hlhu@ieee.org, xuth@sari.ac.cn.

Y. Yang is with the Shanghai Institute of Fog Computing Technology (SHIFT), School of Information Science and Technology, ShanghaiTech University, China, email: yangyang@shanghaitech.edu.cn.

H. Hu is the corresponding author.

Digital Object Identifier: 10.1109/JCN.2018.000036

computing is proposed to improve the service quality of users at the edge of the network. Thus, optimal resource allocation strategies found in existing literatures are developed to maximize users' quality-of-service (QoS). QoS denotes the levels of performance, reliability, and availability [9], such as processing capacity, resource utilization efficiency, processing delay of the cloud data centers.

However, since the actually service quality experienced by the users cannot always be reflected with QoS, quality of experience (QoE), perceived by the end-user, has been introduced to become one of the main guiding paradigms for managing quality provisioning of the cloud and fog computing networks [10]. QoE is actually an extension of QoS by focusing more on the interactivity of the cloud service. Based on the concept of QoE, the authors in [11] measure the average service response-time as the QoE, which is influenced by the queueing delay and round-trip workload transmission time, and then develop workload allocation strategies to maximize the QoE of users under a given power efficiency constraint.

Besides the service quality which is fundamental for users, for cloud providers, the tradeoff between service quality and operational costs needs to be considered. In [12], the authors modeled the resource provisioning problem as an auction-based market for a cloud data center network, where users can develop bidding strategies to compete for the capacity of the cloud data centers with low costs. In fog computing supported medical cyber-physical systems, the authors in [13] jointly study three issues, i.e., BS association, task distribution and VM deployment, to minimize the overall cost while satisfying the QoS requirement, rather than the tradeoff metric including QoS and cost. The authors in [14] focus on the effective and practical resource allocation problem in fog computing. A dynamic resource allocation strategy for fog computing was proposed to improve the efficiency of fog resources utilization and satisfy the users' QoS requirements. By using the proposed strategy, the user can select the satisfying resources autonomously from a group of pre-allocated resources according to the price cost and time cost of task as well as the credibility evaluation of both users and fog resources.

The problem in [15], that how to allocate the limited computing resources of fog nodes to all the users to achieve an optimal and stable performance, is studied and a joint optimization framework was proposed to achieve the optimal resource allocation schemes in a distributed fashion. A Stackelberg game and a many-to-many matching game is applied to investigate the problem and the proposed framework significantly improve the performance of the IoT-based network systems. In this work, the authors maximizing the utilities of users, fog nodes, and cloud data centers, respectively. The authors in [16] analyze the resource management problem in multi-tier data center networks. After modeling the network architecture with 3-layer model, they proposed a hierarchical game, where the interaction between fog nodes and cloud data centers is regarded as a multi-leader multi-follower Stackelberg game, and the interactions between cloud data centers and user are regarded as the single-leader single-follower Stackelberg games. By making decisions distributively, all fog nodes, cloud data centers, and users receive high utilities.

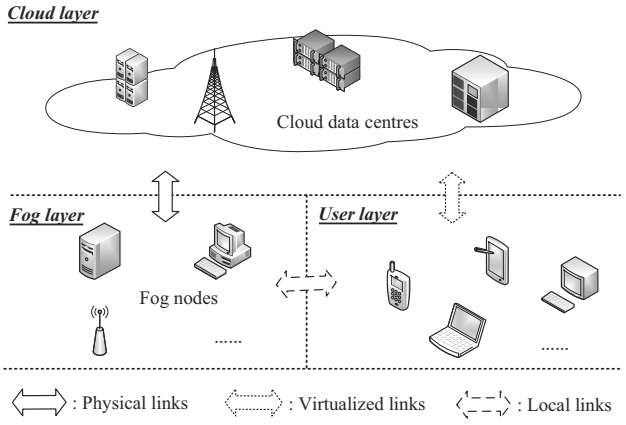

Fig. 1. Three-layer fog computing networks.

In addition, [17] studies the optimal offloading problem in fog computing system to minimize the energy consumption and delay performance. By reconfiguring the offloading probability and varying the transmit power, the objective is to conserve energy for the mobile device and minimize the service delay. The authors in [18] discuss a joint radio and computation resource allocation with the consideration of user energy and delay requirements, and propose a distributive scheme to solve the optimization problem. However, this scenario only consider one centralized cloud provider.

All the above references study the resource allocation problem in fog computing networks. They seek to the optimal QoS, QoE or utility. However, both the service quality and the cost are important metrics for resource allocation. Since improving the service quality tends to increase the cost and the optimal resource allocation solutions based on the two metrics cannot be achieved at the same time, there exists a tradeoff between service quality and cost exacerbated by the presence of service quality targets and economical penalties associated to service quality targets, and the tradeoff between these two metrics needs to be considered.

In our paper, we focus on the cost efficiency including the service quality and overall cost with resource allocation strategy. We first define the concept of cost efficiency in three-layer fog computing networks and then investigate the resource allocation strategy. After formulating the resource allocation problem as a double two-sided matching problem, we proposed a DA-based double-matching strategy (DA-DMS) based on the popular deferred acceptance (DA) algorithm, which is the basic to solve the two-sided matching problem in various systems, such as the coexistence between LTE-U and WiFi [19], [20] and mmWave [21], [22].

III. SYSTEM MODEL

A generic three-layer fog computing network is illustrated as Fig. 1, comprising cloud layer, fog layer and user layer. User layer contains much user equipment that desire low-latency high QoE computing services. Cloud layer consists of large-scale cloud data centers (CDCs) with high-performance computing

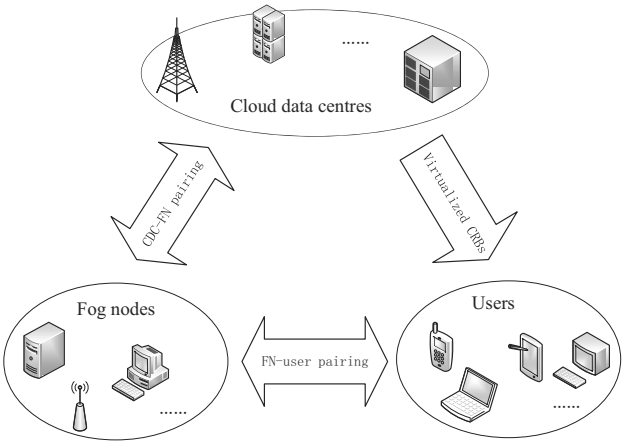

Fig. 2. Interactions among CDCs, FNs, and users.

units which usually located in remote area that can be far from users. Fog layer includes a set of fog nodes (FNs) which are widely deployed closer to the users. FNs can be deployed by the CDCs or other service providers.

FNs are located in the vicinity of users, and hence users can submit their workload to FNs closed to them and the service coverage area can be expended. Furthermore, since all the FNs have limited computing and power resource, the amount of workload submitted from users need to be allocated efficiently. Neighboring FNs can cooperative with each other through local communication infrastructure, and hence they can help each other to jointly process the workload to further improve the service quality of users.

We consider a three-layer fog computing network where each user can submit computing service to a set of neighboring FNs deployed by a set of CDCs. We assume that there are total M CDCs, labeled as $C = \{c_1, c_2, \cdots, c_M\}$. In the virtualized network, all the CDCs serve N users, labeled as $U = \{u_1, u_2, \cdots, u_N\}$. Furthermore, in the physical network, in order to reduce the service delay and achieve real-time fast-response interaction, each CDC can offload its service submitted by users to the local FNs. We assume that there are K FNs in the network, labeled as $F = \{f_1, f_2, ..., f_K\}$.

Let λ_n be the workload arrival rate of user λ_n, $\forall n \in \{1, 2, \cdots, N\}$, and define the unit amount of computing resources that can be distributed by each FN as computing resource block (CRB), which provides computing service at the rate of μ. However, since the users cannot access the CRBs directly, the users are required to receive the virtualized services from the CDCs, and with the management of CDCs, the CRBs of the FNs can finally be allocated to the users. The physical data transmission network between FNs and users satisfies the SecondNet topology [23], where the network resources can be guaranteed for the data center services.

As illustrated in Fig. 2, the interactions among CDCs, FNs, and users can be described as follows.

1. *Interaction between CDCs and users.* In fog computing networks, each users can only access to the CDCs in a virtualized link, and the CDCs provide virtualized CRBs for the users. Different users have various tolerance of service delay, which affects the decision for the number of CRBs. In addition, the cost of the provided CRBs set by CDCs is another key metric for a higher utility.
2. *Interaction between CDCs and FNs.* In addition, the FNs are only accessible to the CDC, too. After providing virtualized CRBs to the users, the CDCs needs to pair FNs distributedly and try to offload the services of users to the FNs by renting the CRBs from the FNs in a physical link. The cost of the rented CRBs from FNs is the critical criterion for CDCs, and a reasonable cost is necessary to maximum utility.
3. *Interaction between FNs and users.* The distance between each FN and each user is different, which will affects the transmission cost. Therefore, within one CDC, the pairing between FNs and users needs to be considered.

We assume that different CDCs offer data services with different requirements and each user has a preference list over all CDCs, denoted as

$$\mathbf{L}_n^{user} = [L_{c1}, L_{c2}, \cdots, L_{cM}], \quad \forall n \in \{1, 2, \cdots, N\}, \tag{1}$$

and hence the each user is required to subscribe to at most one CDC.

In addition, according to the services prices set by all FNs, each CDC has a preference list on each FN while each FN also has different preference on all CDCs with different relations and trading history, i.e.,

$$\mathbf{L}_m^{CDC} = [L_{f1}, L_{f2}, \cdots, L_{fK}], \quad \forall m \in \{1, 2, \cdots, M\}, \tag{2}$$

and

$$\mathbf{L}_k^{FN} = [L_{c1}, L_{c2}, \cdots, L_{cM}], \quad \forall k \in \{1, 2, \cdots, K\}. \tag{3}$$

Furthermore, according to the transmission distance, each FN also has a preference list over all users while each user have preference over FNs based on the rent, i.e.,

$$\mathbf{L}_k^{FN} = [L_{u1}, L_{u2}, \cdots, L_{uN}], \quad \forall k \in \{1, 2, \cdots, K\}, \tag{4}$$

and

$$\mathbf{L}_n^{user} = [L_{f1}, L_{f2}, \cdots, L_{fK}], \quad \forall n \in \{1, 2, \cdots, N\}. \tag{5}$$

According to the preferences, each user purchases the optimal number of virtualized CRBs from the CDCs first. Then, CDCs and FNs needs to be paired to decide the CRBs for users. The most important step is achieving an FN-user pairing result with optimal number of CRBs.

IV. COST EFFICIENCY IN FOG COMPUTING NETWORKS

Before considering the interaction in fog computing and the resource allocation problem, we desire the metric which can be used in the preferences production and the evaluation of resource allocation strategy. Existing works about the resource allocation problem in fog computing are to maximize the service quality [9]–[11] or minimize the cost [12], [14], [15]. In this paper, we consider the two critical metrics simultaneously. Different from the concept of cost efficiency in [13] which minimizes

the overall cost while satisfying the QoS requirement, we define the cost efficiency metric of the network by using the ratio of utility to the overall cost, given as

$$\eta^{CE} = \frac{Utility\ of\ the\ Service\ Quality}{Cost\ of\ the\ Service\ Quality}. \quad (6)$$

In this paper, the service quality can be measured by the service delay including round-trip workload transmission time in the network and the queueing delay at the fog layer. Thus, the cost of service dalay for user u_n, consists of the cost incurred by the queuing delay C_n^{QD} and the cost incurred by the network delay C_n^{ND}, can be measured as

$$C_n^D = C_n^{QD} + C_n^{ND}, \quad \forall n \in \{1, 2, \cdots, N\}. \quad (7)$$

Poisson arrival process is assumed for the workload of each user. Based on the M/G/1 queuing delay model, we adopt the queuing delay cost based on the research in [24], [25]. To model the queueing delay, let $f_n(q_n, \lambda_n)$ denote the queueing delay at user u_i given q_n active servers and an arrival rate of λ_n. If the workload is perfectly parallelizable, and arrivals are Poisson, then $f_n(q_n, \lambda_n)$ is the average delay of q_n parallel queues, each with arrival rate λ_n/q_n. Moreover, if each queue is an M/G/1 Processor Sharing (PS) queue, then $f_n(q_n, \lambda_n) = 1/(\mu - \lambda_n/q_n)$. In this fog computing networks, the queueing delay is equal to the mean response time for unit transmitted data as $\bar{t} = [1/(\mu - (\lambda_n/q_n))]$. Thus, the cost of queuing delay for user u_n can be given as

$$C_n^{QD} = \bar{t} \cdot \lambda_n = \frac{\lambda_n}{\mu - \lambda_n/q_n}, \quad \forall n \in \{1, 2, \cdots, N\}, \quad (8)$$

where q_n is the number of CRBs purchased by user u_n.

In addition, the network delay, i.e., C_n^{ND}, is related to the transmission distance, traffic condition and many other unpredicted factors. Since the networking delay model is uncorrelated to the working of the matching problem which we focus on in this paper, we assume the network resource from the FNs to users is always sufficient and adopt a simple ideal model of the network delay as did in [15], where the cost incurred by the network delay C_n^{ND} is a linear function of l_{kn} for simplicity, i.e.,

$$C_n^{ND} = \alpha \cdot l_{kn}, \quad \forall n \in \{1, 2, \cdots, N\}, \quad (9)$$

where α is a scalar, and l_{kn} denotes the distance from the CDC to the FN plus the distance from the FN f_k to the user u_n.

Therefore, the cost of service dalay for user u_n based on (7) can be rewritten as

$$C_n^D = \frac{\lambda_n}{\mu - \lambda_n/q_n} + \alpha \cdot l_{kn}, \quad \forall n \in \{1, 2, \cdots, N\}. \quad (10)$$

A. Cost Efficiency for CDCs

In fog computing networks, the utility of each CDC is the revenue received from the users minus the payment to the facilities that are able to provide the physical CRBs.

By CDC-FN pairing, each CDC prefers to offload its services to the FNs nearby based on the number of virtualized CRB purchased by serving users. However, if there are not sufficient available CRBs from the FNs which can satifity the requirements of all users. Some users will still be served by the remote CDC located far away from the users, and hence pay more the energy cost. Therefore, the utility function of the CDC $c_m, m \in \{1, 2, \cdots, M\}$ can be given as

$$U_m^{CDC} = \sum_{n=1}^{N} \tau_{nm} p_m^{CDC} q_n - \sum_{k=1}^{K} p_k^{FN} q_{nk}^{FN} - e_i q_n^{CDC}, \quad (11)$$

where τ_{nm} is set to be the Boolean variable determining whether CDC c_m serves user u_n or not. p_m^{CDC} denotes the price set by CDC c_m for each unit of the virtualized CRB, and hence $\sum_{n=1}^{N} \tau_{nm} p_m^{CDC} q_n$ denote the total revenue that CDC c_m receives from users for its data services. $\sum_{k=1}^{K} p_k^{FN} q_{nk}^{FN}$ denotes the total payment from the CDC c_m to all FNs. p_k^{FN} denotes the price set by the FN f_k, which is determined by the cost and current traffic. q_{nk}^{FN} CRBs are offloaded to the FN f_k, and q_n^{CDC} CRBs are still served by the remote CDC. e_i presents the increment of the energy cost in the remote CDC.

According to the cost in (10) and the utility in (11), the cost efficiency for CDC c_m can be given as

$$\eta_n^{CDC} = \frac{U_m^{CDC}}{\sum_{n=1}^{N} \tau_{nm} C_n^D}, \quad \forall n \in \{1, 2, \cdots, N\}. \quad (12)$$

B. Cost Efficiency for FNs

For FN f_k in the fog computing network, the utility is the payment received from the CDC minus the transmission cost which can be given as

$$U_k^{FN} = \sum_{n=1}^{N} (p_k^{FN} - o_{nk}) q_{nk}^{FN}, \quad \forall k \in \{1, 2, \cdots, K\}, \quad (13)$$

where o_{nk} denotes the payment for the transmission of each unit CRB, which has a linear relationship with the distance l_{nk}.

Therefore, the cost efficiency for FN f_k can be given as

$$\eta_k^{FN} = \frac{U_k^{FN}}{\sum_{n=1}^{N} \tau_{nk} C_n^D}, \quad \forall k \in \{1, 2, \cdots, K\}. \quad (14)$$

C. Cost Efficiency for Users

Users in the fog computing network pay CDCs for the service, and the utility of user, $u_n, \forall n \in \{1, 2, \cdots, N\}$, can be denoted as the revenue received from the workload data minus both the cost of service delay and payment to the CDCs, which can be expressed as

$$U_n^{USER} = \sum_{m=1}^{M} \tau_{nm} (\alpha_n \lambda_n - \beta_n q_n p_m^{CDC} - \gamma_n C_n^D), \quad (15)$$

where α_n denotes the revenue that user u_n can obtain for unit received data rate. β_n and β_n present the weight factors indicating the importance of payment and data service delay in the utility function of user u_n, respectively.

Thus, the cost efficiency for user u_n can be written as

$$\eta_n^{USER} = \frac{U_n^{USER}}{C_n^D}, \quad \forall n \in \{1, 2, \cdots, N\}. \quad (16)$$

As described in (11), (14) and (16), the cost efficiency for CDCs, FNs and users can be used for resource allocation problem in fog computing networks.

V. RESOURCE ALLOCATION STRATEGY IN FOG COMPUTING NETWORKS

A. Double-Matching Resource Allocation Problem

We assume that each user has purchased the optimal number of virtualized CRBs from the CDCs. According to the preferences based on cost efficiency, CDC-FN and FN-user need to be paired. These two pairing problems can be formulated in a unified form as a binary-integer programming problem based on the system metric maximization.

Let $x_{i,j}$ be a binary variable. $x_{i,j} = 1$, if the *j*th object on the side is paired with the *i*th object on the other side, otherwise, $x_{i,j} = 0$. Mathematically, the pairing problem can be formulated as

$$\max \quad \sum_{i=1}^{I}\sum_{j=1}^{J} x_{i,j}\eta_{i,j}, \tag{17a}$$

$$\text{s.t.} \quad \sum_{j=1}^{J} x_{i,j} = 1, \quad \forall i, \tag{17b}$$

$$\sum_{i=1}^{I} x_{i,j} = 1, \quad \forall j, \tag{17c}$$

where constraints (17b) and (17c) indicate that each object on the side can be only allocated to one object on the other side.

In the formulated problem, the cost efficiency is adopted as the metric, i.e., $\eta_{i,j}$. With the consideration of double two-sided matching, we need to get the cost efficiency in CDC-FN pairing and FN-user pairing, respectively. In CDC-FN paring, the cost efficiency is defined by the ratio between the utility of CDC and FN and the cost, and each user served by (m, k) CDC-FN pair should be considered.Thus, the cost efficiency can be give as

$$\eta_{i,j}|_{CDC-FN} = \eta_{m,k} = \sum_{n=1}^{N} \frac{U_m^{CDC}|_n + U_k^{FN}}{\tau_{nm}\tau_{nk}C_n^D}, \tag{18}$$

and the cost efficiency of FN-user pairing can be defined by the ratio between the utility of FN and user and the cost as

$$\eta_{i,j}|_{FN-user} = \eta_{k,n} = \frac{U_k^{FN} + U_n^{USER}}{\sum_{i=1}^{N}\tau_{ik}C_i^D}, \tag{19}$$

where the utility of CDC, FN and user can be given by (11), (13) and (15), respectively.

The problem form in (17) which include two pairing problems is a classical integer linear programming problem, which is NP-hard. However, in three-layer fog computing networks, the two pairing problem cannot be considered independently, since the selected CDC will affect the pairing between FN and user. Thus, we reformulate the two pairing problem as a unitary double-matching optimization problem as

$$\max \quad \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{n=1}^{N} x_{m,k,n}\cdot\eta_{m,k,n}, \tag{20a}$$

$$s.t. \quad \sum_{m=1}^{M}\sum_{k=1}^{K} x_{m,k,n} = 1, \quad \forall n, \tag{20b}$$

where the binary variable, $x_{m,k,n} = 1$, if the CDC c_m serve user u_n by offloading the service of user u_n to the FN f_k, otherwise, $x_{m,k,n} = 0$. Here, we assume that the CRBs of one user can be supported by only one CDC and only one corresponding FN while both each CDC and FN can serve multiple users. Therefore, the constraint (20b) implies that each user can be only assigned to one CDC and one FN. The unitary metric, $\eta_{m,k,n}$, can be given based the description above as

$$\eta_{m,k,n} = \frac{U_{m,k,n}^{\text{total}}}{C_n^D}, \tag{21}$$

where $U_{m,k,n}^{\text{total}}$, given in (22), denotes the total utility when c_m, f_k, and u_n are paired to allocate resources. $U_m^{CDC}|_{k,n}$ denotes the utility of CDC c_m for FN f_k and user u_n. $U_k^{FN}|_n$ denotes the utility of FN f_k for user u_n. $U_n^{USER}|_m$ denotes the utility of user u_n for CDC c_m. Thus, the double-matching problem is more complex than problem (17) and is also an NP-hard problem.

Next, we will study this problem from a novel perspective and develop an efficient heuristic algorithm. Actually, the formulated double-matching problem is a two-step matching problem as illustrated in Fig. 3, where each user has purchased the optimal number of virtualized CRBs from the CDCs, and the purchased CRBs needs to be allocated to CDCs and FNs. Thus, for each CRB of all user, one CDC and one FN need to be selected. Therefore, based on the cost efficiency, each user is matched with one CDC and one FN.

Furthermore, the matching is based the preferences among CDCs, FNs and users. The preferences is calculated according to the cost efficiency as given in (21). As shown in Fig. 3, each CDC has a preference list for all FNs, and the green line of dashes denotes the preferred FN for each CDC. Similarly, each FN has a preference for all CDCs, and FNs and users have the preference mutually. The double-matching resource allocation is to find optimal CDC-FN-user pairs for total users to maximize the whole cost efficiency as far as possible.

B. Double-Matching Resource Allocation Strategy

In this paper, we proposed a double-matching strategy based on deferred acceptance (DA-DMS) based on the DA algorithm

$$\begin{aligned} U_{m,k,n}^{\text{total}} &= U_m^{CDC}|_{k,n} + U_k^{FN}|_n + U_n^{USER}|_m \\ &= \tau_{nm}p_m^{CDC}q_n - p_k^{FN}q_{nk}^{FN} - e_i q_n^{CDC} + (p_k^{FN} - o_{nk})q_{nk}^{FN} + \tau_{nm}(\alpha_n\lambda_n - \beta_n q_n p_m^{CDC} - \gamma_n C_n^D) \\ &= \tau_{nm}[(1-\beta_n)p_m^{CDC}q_n + \alpha_n\lambda_n - \gamma_n C_n^D] - e_i q_n^{CDC} - o_{nk}q_{nk}^{FN}. \end{aligned} \tag{22}$$

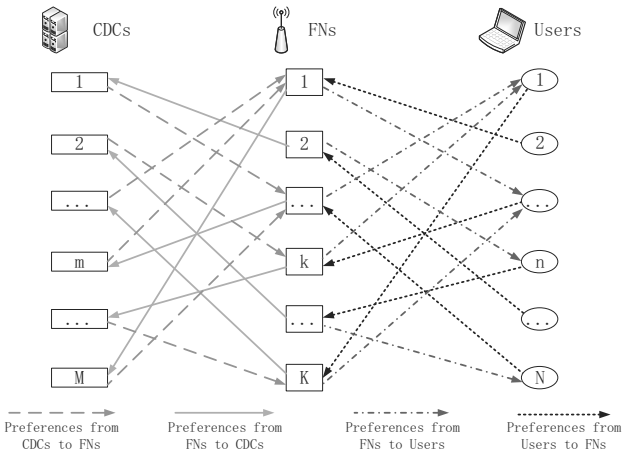

Fig. 3. Double-matching problem in fog computing networks.

which is designed for the common many-to-one matching problem.

The traditional DA algorithm was designed in [26] for the college admission problem, which is a one-step matching problem. Due to the double-matching, the DA algorithm cannot be directly applied to the resource allocation problem in fog computing networks. In the the double-matching resource allocation problem, CDCs and FNs are coupled for users while FNs and users are coupled for CDCs. In other words, CDC-FN pairing needs to consider the influence between FNs and users and FN-user pairing needs to consider the constraints between CDCs and FNs.

Therefore, to take these constraints into consideration, we develop the DA-DMS strategy for the double-matching resource allocation problem based on the DA algorithm. The DA-DMS strategy for the double-matching resource allocation problem is described in Algorithm 1. This proposed strategy considers CDC-FN pairing first based on the DA algorithm, and begins from FN in the CDC-FN pairing procedure while begins from FN in the FN-user pairing procedure. In other words, both the CDC-FN pairing and the FN-user pairing is FN-propose .

In two-sided matching, Gale and Shapley [26] have proved that there always exists one-side optimal stable matching with deferred acceptance algorithm. For example, man-woman marriage matching, when all men and women have strict preferences, there always exists an M-optimal stable matching (that every man likes at least as well as any other stable matching), and a W-optimal stable matching. Furthermore, the matching produced by the deferred acceptance algorithm with men proposing is the M-optimal stable matching. The W-optimal stable matching is the matching produced by the algorithm when the women propose. In this paper, both the CDC-FN matching and FN-user matching are treated when FNs propose, and hence these two matchings are FN-optimal stable matchings. In addition, the proposed DA-DMS does the CDC-FN matching and FN-user matching iteratively. Thus, the proposed DA-DMS is FN-optimal stable matching.

A simple example of the proposed DA-DMS strategy is illustrated in Fig. 4. We can find that the final results are different with the results after step 2. In Fig. 4, there are 6 CDCs, 6 FNs and 6 users. Based on the cost efficiency for CDC-FN pairs as described in (18), the performance metric of each CDC to each FN can be given as a 6×6 matrix, taking an example for sim-

Algorithm 1 Double two-sided matching strategy based on deferred acceptance algorithm (DA-DMS)

Require: C, F, U and total parameters.
Ensure: *CDC-FN Pairing (CDC-[FN-user] Pairing.)*
1: Calculate the cost efficiency for CDC-FN pairs by recursively iterating total candidates as (18), i.e., $\eta_{m,k}$, $\forall m \in \{1,2,\cdots,M\}, \forall k \in \{1,2,\cdots,K\}$.
2: $\forall k \in \{1,2,\cdots,K\}$, $\alpha(k) = \emptyset$, the waiting list $\mathbf{L}_k^{wait} = \emptyset$; $\forall m \in \{1,2,\cdots,M\}$, the candidate list $\mathbf{L}_m^{cand} = \{1,2,\cdots,K\}$.
3: **while** $\exists$ FNs with empty waiting list **do**
4: $k'_m :=$ the most-preferred FN in candidate list $\mathbf{L}_m^{cand}$ based on $\eta_{m,k}$.
5: **for** k=1:K **do**
6: **for** m=1:M **do**
7: **if** $k'_m == k$ **then**
8: **if** $\mathbf{L}_k^{cand} == \emptyset$ **then**
9: $\mathbf{L}_k^{wait} = \mathbf{L}_k^{wait} \bigcup k'_m$.
10: **else**
11: $\mathbf{L}_m^{cand} = \mathbf{L}_m^{cand} \backslash \{k\}$
12: **end if**
13: **end if**
14: **end for**
15: **end for**
16: **end while**
17: $\forall k \in \{1,2,\cdots,K\}$, $\alpha(k) = \mathbf{L}_k^{wait}$.
Ensure: *[CDC-FN]-user Pairing*
18: Calculate the cost efficiency for each CDC-FN-user pair by recursively iterating total candidates as (21), i.e., $\eta_{m,k,n}$, $\forall m \in \{1,2,\cdots,M\}$, $\forall k \in \{1,2,\cdots,K\}$, $\forall n \in \{1,2,\cdots,N\}$.
19: $\forall n \in \{1,2,\cdots,N\}$, $\beta(n) = \emptyset$, the waiting list $\mathbf{L}_n^{wait} = \emptyset$; $\forall k \in \{1,2,\cdots,K\}$, the candidate list $\mathbf{L}_k^{cand} = \{1,2,\cdots,N\}$.
20: **while** $\exists$ users with empty waiting list **do**
21: $n'_k :=$ the most-preferred user in candidate list $\mathbf{L}_k^{cand}$ based on $\eta_{\alpha(k),k,n}$.
22: **for** n=1:N **do**
23: **for** k=1:K **do**
24: **if** $n'_k = n$ **then**
25: **if** $\mathbf{L}_k^{cand} = \emptyset$ **then**
26: $\mathbf{L}_n^{wait} = \mathbf{L}_n^{wait} \bigcup n'_k$.
27: **else**
28: $\mathbf{L}_k^{cand} = \mathbf{L}_k^{cand} \backslash \{n\}$.
29: **end if**
30: **end if**
31: **end for**
32: **end for**
33: **end while**
34: $\forall n \in \{1,2,\cdots,N\}$, $\beta(n) = \mathbf{L}_n^{wait}$.
35: Do CDC-[FN-user] pairing and [CDC-FN]-user pairing recursively, until (α, β) is changeless.
36: **return** α, β.

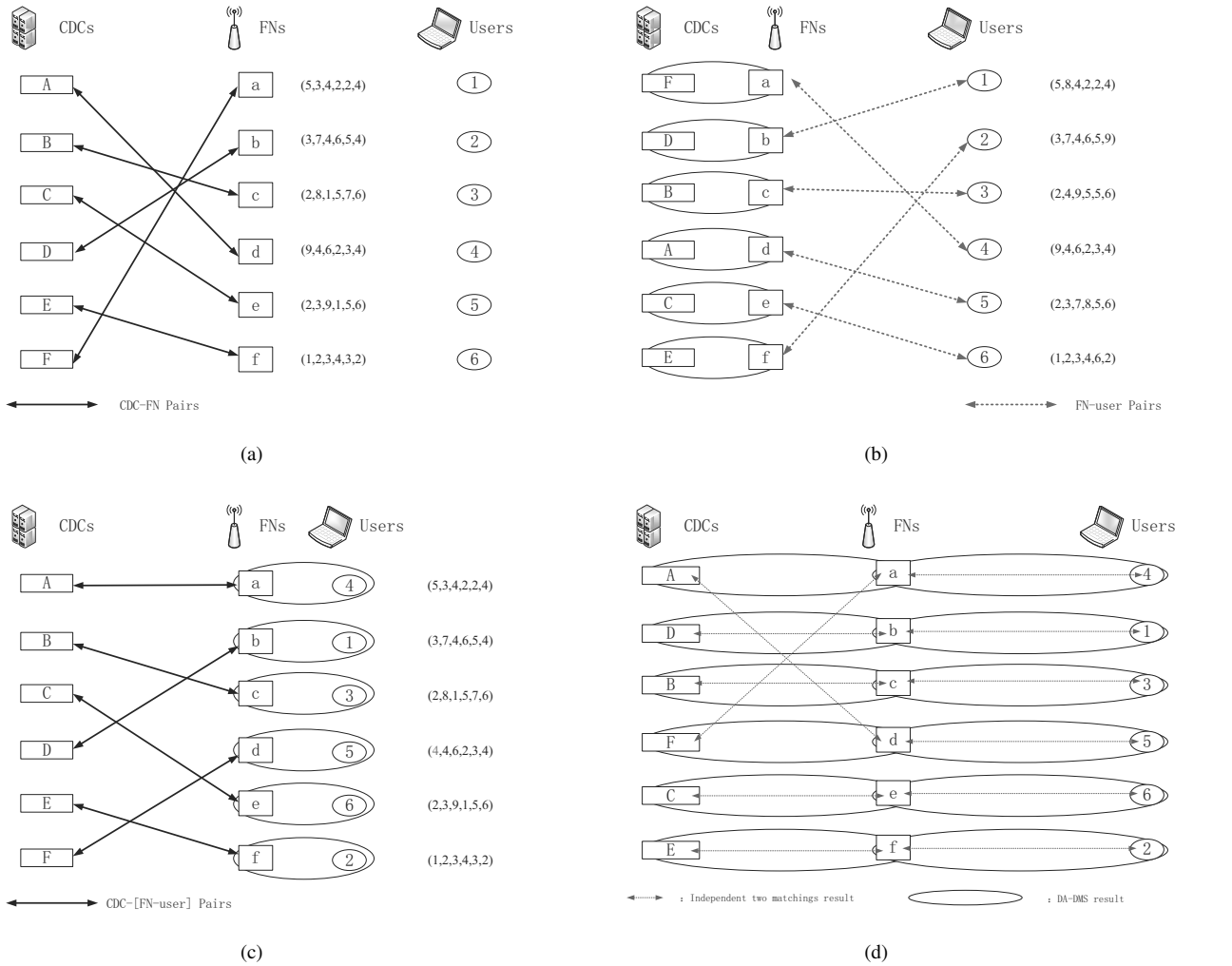

Fig. 4. Simple example for the proposed DA-DMS strategy: (a) CDC-FN pairing, (b) [CDC-FN]-user pairing, (c) CDC-[FN-user] pairing, and (d) final pairing result.

plicity, as follows

$$\eta_{m,k} = \begin{bmatrix} 5 & 3 & 4 & 2 & 2 & 4 \\ 3 & 7 & 4 & 6 & 5 & 4 \\ 2 & 8 & 1 & 5 & 7 & 6 \\ 9 & 4 & 6 & 2 & 3 & 4 \\ 2 & 3 & 9 & 1 & 5 & 6 \\ 1 & 2 & 3 & 4 & 3 & 2 \end{bmatrix}. \tag{23}$$

The proposed strategy mainly consists of CDC-FN pairing procedure and [CDC-FN]-user pairing procedure. The CDC-FN pairing procedure is the matching between CDC and FN with consideration of the cost efficiency and the CRB allocation, while the [CDC-FN]-user pairing procedure is based on the results of the CDC-FN pairing and the aim is to find the optimal FN-user pairs. However, such an independent matching procedures ignore the interaction among CDCs, FNs and users. Thus, recursively iterations is necessary to repair CDCs and FNs when FN-user pairing has been processed.

According to the proposed strategy as described in Algorithm 1, in the CDC-FN pairing procedure as illustrated in Fig. 4(a), the CDC-FN pairing results are {(F-a),(D-b),(B-C),(A-D),(C-e),(E-f)}. Furthermore, based on the cost efficiency for [CDC-FN]-user pairs, i.e., $\eta_{\alpha(k),k,n}$, the performance metrics between each FN and each user can be given as a matrix. Here, in the simple example, the matrix is

$$\eta_{\alpha(k),k,n} = \begin{bmatrix} 5 & 8 & 4 & 2 & 2 & 4 \\ 3 & 7 & 4 & 6 & 5 & 9 \\ 2 & 4 & 9 & 5 & 5 & 6 \\ 9 & 4 & 6 & 2 & 3 & 4 \\ 2 & 3 & 7 & 8 & 5 & 6 \\ 1 & 2 & 3 & 4 & 6 & 2 \end{bmatrix}, \tag{24}$$

and based on the cost efficiency for CDC-[FN-user] pairs, i.e., $\eta_{m,\beta(n),n}$, the performance metrics between each CDC and each FN after FN-user pairing can be given as

$$\eta_{m,\beta(n),n} = \begin{bmatrix} 5 & 3 & 4 & 2 & 2 & 4 \\ 3 & 7 & 4 & 6 & 5 & 4 \\ 2 & 8 & 1 & 5 & 7 & 6 \\ 4 & 4 & 6 & 2 & 3 & 4 \\ 2 & 3 & 9 & 1 & 5 & 6 \\ 1 & 2 & 3 & 4 & 3 & 2 \end{bmatrix}, \tag{25}$$

where we just set the performance metric as same as the performance metric of each CDC to each FN without FN-user pairing

Table 1. Parameters and values used for analysis.

Parameter	Symbol	Value
Number of CDCs	M	4
Number of FNs	K	20
Number of users	N	40:20:200
Service rate of each CRB	μ	$0.1ms^{-1}$
Workload arriving rate of each user	λ	$0.5ms^{-1}$
Number of CRBs purchased by each user	q	U(0,100)

besides one different entry of the matrix. As illustrated in Fig. 4, we can find that the final results is different from that of independent two matchings. In Fig. 4 (d), the we can find the difference between the proposed DA-DMS algorithm and independent two matchings is the FN selection of CDC A and F, and according to the value of metric in the example, we can find that the proposed algorithm can improve the cost efficiency from 84 to 88.

In the CDC-FN pairing procedure of the developed strategy, α denotes the pairing result that FN f_k is matched with FN $c_{\alpha(k)}$. While β is the pairing result in the FN-user pairing procedure when CDC and FN have been matched, i.e., user u_n is paired with FN $f_{\beta(n)}$. By recursively iterating the two matching procedures, the final double-matching result can be given based on α and β.

VI. NUMERICAL RESULTS

In this section, we consider fog computing IoT network scenario to evaluate the performance and present simulation results with MATLAB. In the simulated scenario, for simplicity, we conFig. 4 CDCs, 20 FNs, and 120 users allocated randomly in a circle district where each user and its sensor are located at the same position. In this systems, users are mobile devices with computational intensive applications, and the fog nodes are the access points (e.g., cellular base stations or WLAN access points.) managed by a network operator. The detailed simulation parameter settings are listed in Table 1. We set the service rate of each CRB is $0.1\ \mathrm{ms}^{-1}$, and the workload arriving rate of each user is a random number averaged $0.5\ \mathrm{ms}^{-1}$. Furthermore, for each FN, we set its preference to each CDC based on the cost efficiency, and the amount of available CRB as a random number satisfying uniform distribution between $(0, 100)$.

Fig. 5 presents the cost efficiency of the whole system with the number of users increasing. Here, we compare the cost efficiency performance with various workload. With the with same workload, we can find that the cost efficiency increases generally when the number of users increases, and the cost efficiency performance achieve a higher value when the average workload for each user increases.

Furthermore, as shown in Fig. 6, we evaluate the cost efficiency performance when the number of users increases with various number of FNs. When the number of FNs is fixed, the cost efficiency performance generally increases with the increasing of the user number, and moreover, the increasing speed first increases then gradually decreases to zero. It is because when the number of users increases, the FNs can serve more active users. Furthermore, the user will be served from CDCs when all of the available CRBs of FNs are allocated and the number of user keep increasing. Thus, the cost efficiency performance of the FNs stop increasing. In addition, we can find that when the

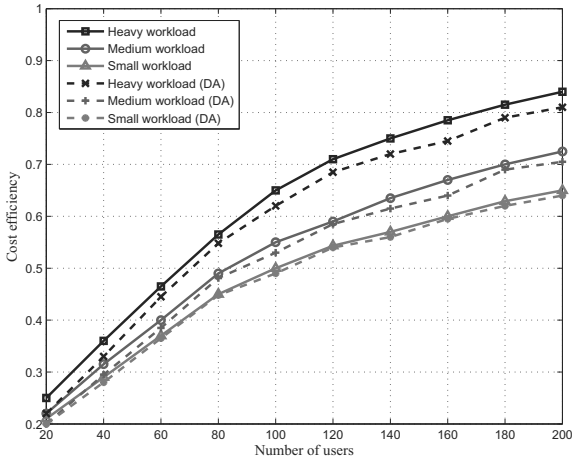

Fig. 5. Cost efficiency versus the number of users with various workload levels.

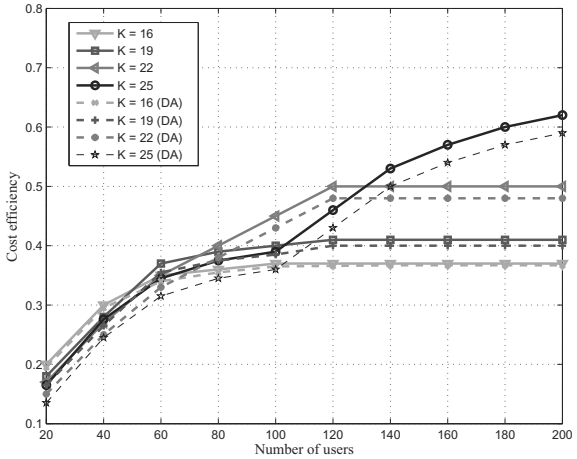

Fig. 6. Cost efficiency versus the number of users various number of FNs.

number of users is large, more FNs can help to converge more users, while FNs will compete when small number of user, and hence the increasing speed becomes small.

In Fig. 7, we evaluate the cost efficiency performance with various service rate of each CRB to find the relationship between the cost efficiency performance and average service rate of each CRB. As shown in the figure, we can find the cost efficiency performance increases when the value of service rate increases. The reason is that when service rate increases, each user is able to be served with a less number of CRBs. Therefore, the CDC can set a higher price to the users and receive high cost efficiency. At the same time, since CDCs can serve more users when the number of users increases, cost efficiency performance increases.

In addition, the proposed DA-DMS algorithm is better than the traditional DA algorithm for the formulated double two-sided matching problem obviously as present in Fig. 4. The same results can be found in the performance comparisons as present in Figs. 5–7, i.e., the DA-DMS algorithm is better than the tradi-

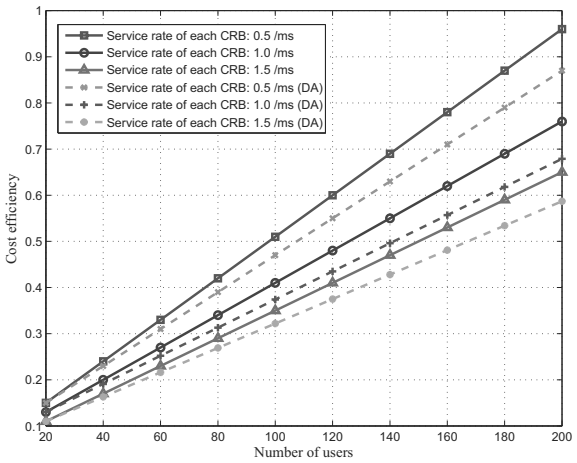

Fig. 7. Cost efficiency versus the number of users with various service rate.

tional DA algorithm for the double two-sided matching problem, especially in complex scenarios (heavy workload, more number of FNs).

VII. CONCLUSION

In this paper, we investigated the resource allocation problem in three layers fog computing networks. We first formulated the problem as a double-matching problem, and proposed the definition of cost efficiency which can be used in the preference analysis among CDCs, FNs and users. Then, based on the cost efficiency, we developed a double-matching strategy based on deferred acceptance algorithm (DA-DMS). By using the DA-DMS strategy, the three participants could achieve stable results that each participant cannot change its paired partner unilaterally for more cost efficiency. Numerical results showed that high cost efficiency performance could be achieved by adopting the DA-DMS strategy. In addition, the double-matching problem can be extended to further three-layer networks, such as unmanned aerial vehicle [27], [28], which will be researched in our future work.

REFERENCES

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

[2] J. Lin *et al.,* "A Survey on Internet of Things: architecture, enabling technologies, security and privacy, and applications," *IEEE Internet Things J.,* vol. 4, no. 5, pp. 1125–1142, Oct. 2017.

[3] D. Hoang, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches," *Wireless Commun. Mobile Comput.,* vol. 13, no. 18, pp. 1587–1611, Dec. 2013.

[4] O. Osanaiye *et al.,* "From cloud to fog computing: A review and a conceptual live vm migration framework," *IEEE Access,* vol. 5, pp. 8284–8300, 2017.

[5] S. Yi, Z. Hao, Z. Qin, and Q. Li, "Fog computing: Platform and applications," in *Proc. IEEE HotWeb,*, Nov. 2015, pp. 73–78.

[6] 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.

[7] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proc. MCC Workshop Mobile Cloud Comput.,* 2012, pp. 13–16.

[8] C. Mouradian *et al.,* "A comprehensive survey on fog computing: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.,* vol. 20, no. 1, pp. 416–464, Firstquarter 2018.

[9] D. Ardagna, G. Casale, M. Ciavotta, J. F. Pérez, and W. Wang, "Quality-of-service in cloud computing: Modeling techniques and their applications," *J. Internet Serv. Appl.,* vol. 5, no. 1, pp. 11, 2014.

[10] P. Casas and R. Schatz, "Quality of experience in cloud services: Survey and measurements," ,*Comput. Netw.,* vol. 68, pp. 149-165, 2014.

[11] Y. Xiao and M. Krunz, "QoE and power efficiency tradeoff for fog computing networks with fog node cooperation," in *Proc. IEEE INFOCOM*, 2017, pp. 1–9.

[12] L. Zheng, C. Joe-Wong, C. W. Tan, M. Chiang, and X. Wang, "How to bid the cloud," *ACM SIGCOMM Comput. Commun. Rev.,* vol. 45, no. 4, pp. 71–84, 2015.

[13] L. Gu, D. Zeng, S. Guo, A. Barnawi, and Y. Xiang, "Cost efficient resource management in fog computing supported medical cyber-physical system," *IEEE Trans. Emerg. Topics Comput.,* vol. 5, no. 1, pp. 108–119, Jan.–Mar. 2017.

[14] L. Ni, J. Zhang, C. Jiang, C. Yan, and K. Yu, "Resource allocation strategy in fog computing based on priced timed petri nets," *IEEE Internet Things J.,* vol. 4, no. 5, pp. 1216–1228, Oct. 2017.

[15] H. Zhang *et al.,* "Computing resource allocation in three-tier IoT fog networks: A joint optimization approach combining stackelberg game and matching," *IEEE Internet Things J.,* vol. 4, no. 5, pp. 1204–1215, Oct. 2017.

[16] H. Zhang *et al.*, "Fog computing in multi-tier data center networks: A hierarchical game approach," in *Proc. IEEE ICC*, May 2016, pp. 1–6.

[17] L. Liu, Z. Chang, X. Guo, S. Mao, and T. Ristaniemi, "Multi-objective optimization for computation offloading in fog computing," *IEEE Internet Things J.,* vol. 5, no. 1, pp. 283–294, Feb. 2018.

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

[19] Y. Gu, C. Jiang, L.X. Cai, M. Pan, L. Song, and Z. Han, "Dynamic path to stability in the LTE-Unlicensed with user mobility: A dynamic matching framework," *IEEE Trans. Wireless Commun.,* vol. 15, no. 7, pp. 4547–4561, July 2017.

[20] Y. Gu, Y. Zhang, L. X. Cai, M. Pan, L. Song, and Z. Han, "LTE-Unlicensed co-existence mechanism: A matching game framework," *IEEE Wireless Commun.* vol.23, no.6, pp.54-60, Dec. 2016.

[21] O. Semiari, W. Saad, S. Valentin, M. Bennis, and H. V. Poor, "Context-aware small cell networks: How social metrics improve wireless resource allocation," *IEEE Trans. Wireless Commun.,* vol. 14, no. 11, pp. 5927–5940, Nov. 2015.

[22] O. Semiari, W. Saad, and M. Bennis, "Joint millimeter wave and microwave resources allocation in cellular networks with dual-mode base stations," *IEEE Trans. Wireless Commun.,* vol. 16, no. 7, pp. 4802-4816, July 2017.

[23] C. Guo *et al.*, "SecondNet: A data center network virtualization architecture with bandwidth guarantees," in *Proc. Int. Conf. ACM,* Nov./Dec. 2010, pp. 1–5.

[24] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Geographical load balancing with renewables," *ACM SIGMETRICS Perform. Eval. Rev.,* vol. 39, no. 3, pp. 62–66, Dec. 2011.

[25] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew. "Greening geographic load balancing," in *Proc. ACM Sigmetrics,* 2011.

[26] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *American Mathematical Monthly,* pp. 9–15, 1962.

[27] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," *IEEE Trans. Wireless Commun.,* vol. 15, no. 6, pp. 3949–3963, June 2016.

[28] S. Jeong, O. Simeone, and J. Kang, "mobile edge computing via an uav-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Trans. Veh. Technol.,* vol. 67, no. 3, pp. 2049–2063, Mar. 2018.

Boqi Jia received the B.S. degree from the Yingcai Honors College, University of Electronic Science and Technology of China, Chengdu, China, in 2013. He is currently pursuing the Ph.D. degree with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. He has won the Best Paper Award of *IEEE GLOBECOM 2016* as the first author. His research interests include V-MIMO systems, user pairing, EE-SE tradeoff, unlicensed spectrum, and 5G mobile systems.

Honglin Hu received his Ph.D. degree in Communications and Information System in 2004, from the University of Science and Technology of China. Then, he was with Future Radio, Siemens AG Communications in Munich, Germany. Since 2009, he served as a Full Professor at Shanghai Institute of Microsystem and Information Technology. Also, he serves as an adjunct professor at ShanghaiTech University and the vice director of Shanghai Research Center for Wireless Communications since 2013. Since 2016, he joined Shanghai Advanced Research Institute (SARI), Chinese Academy of Sciences. Dr. Hu was the Vice Chair of IEEE Shanghai Section and the leading guest editor for *IEEE Wireless Communications* special issue on "Mobile Converged Networks" and *IEEE Communications Magazine* special issue on "Software Defined Wireless Networks (Part I and Part II)". He received 2016 IEEE Jack Neubauer Memorial Award (the best paper award of *IEEE Transactions on Vehicular Technology*) and the best paper award of *IEEE GLOBECOM 2016*. He is also a Finland Distinguished Professor (FiDiPro) at VTT, Finland (2015–2018).

Yu Zeng received the B.S. degree from Beijing University of Posts and Telecommunications, Beijing, China in 2013. He is currently pursuing the Ph.D. degree with the Shanghai Institute of Microsystem and Information technology, Chinese Academy of Sciences, Shanghai, China. He has received the Best Paper Award of *IEEE GLOBECOM 2016*. His research interests include D2D communications, user pairing, unlicensed spectrum, and performance optimization in the fifth-generation networks.

Tianheng Xu received the Ph.D. degree from Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, in 2016. Dr. Xu is currently an Assistant Professor with Shanghai Advanced Research Institute, Chinese Academy of Sciences. His main research interests include 5G wireless communications, statistical signal processing, and vehicular communication technologies. Dr. Xu received the President Scholarship (First Grade) of Chinese Academy of Sciences in 2014, the Outstanding Ph.D. Graduates Award of Shanghai in 2016, and the Shanghai Young Talent Sailing Program in 2017. He also received the Best Paper Award at the *IEEE GLOBECOM 2016*.

Yang Yang received the B.Eng. and M.Eng. degrees in Radio Engineering from Southeast University, Nanjing, China, in 1996 and 1999, respectively; and the Ph.D. degree in Information Engineering from the Chinese University of Hong Kong in 2002. He is currently a Professor with the School of Information Science and Technology, ShanghaiTech University, China, serving as a Co-Director of Shanghai Institute of Fog Computing Technology (SHIFT). His current research interests include wireless sensor networks, Internet of Things, Fog computing, Open 5G, and advanced wireless testbeds. He has published more than 180 papers and filed over 80 technical patents in wireless communications. He is a Fellow of the IEEE.