IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON EDGE INTELLIGENCE FOR INTERNET OF THINGS

Received July 19, 2020, accepted August 11, 2020, date of publication August 20, 2020, date of current version September 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018166

Service Pricing and Selection for IoT Applications Offloading in the Multi-Mobile Edge Computing Systems

WANLI ZHANG[1], XIANWEI LI[1], LIANG ZHAO[2], XIAOYING YANG[1], TAO LIU[3], AND WEI YANG[4]

[1]School of Information Engineering, Suzhou University, Suzhou 234000, China
[2]School of Computer Science, Shenyang Aerospace University, Shenyang 110000, China
[3]School of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China
[4]School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233000, China

Corresponding author: Xianwei Li (lixianwei163@163.com)

This work was supported in part by the Academic and Technical Leaders under Grant 2018XJHB07, in part by the Suzhou Science and Technology Project under Grant SZ2018GG01, and Grant SZ2018GG01xp, in part by the Outstanding Academic and Technical Backbone of Suzhou University under Grant 2016XJGG12, in part by the New Engineering Research and Practice Projects under Grant 2017xgkxm54, in part by the National Science Foundation for Young Scientists of China under Grant 61701322, in part by the Scientific Research Project and the Teaching Research Project of Anhui University of Finance and Economics under Grant acjyyb2019063 and Grant ACKYC19049, in part by the Young and Middle-Aged Science and Technology Innovation Talent Support Plan of Shenyang under Grant RC190026, in part by the Collaborative Education Project under Grant 201902167037 and Grant 201902167019, in part by the Key Curriculum Construction Project under Grant szxy2018zdkc19, in part by the Large Scale Online Open Course (MOOC) Demonstration Project under Grant 2019mooc300 and Grant 2019mooc318, and in part by the Professional Leader of Suzhou University under Grant 2019XJZY22.

ABSTRACT With the rapid developing of Internet of Things (IoT) technologies, various kinds of IoT devices are connected over the Internet. Consequently, how to meet the requirements of executing IoT applications is becoming a critical issue. Offloading the IoT applications to the public cloud is an efficient approach to enhance the computing capabilities of these IoT devices. However, as there is a long distance from the IoT devices to the remote public cloud, transmission delay will be caused. Mobile edge computing (MEC) provides an effective solution to this issue since IoT devices are near to the servers of the MEC systems. Pricing and load balancing are two important factors for cloud service provision. Pricing is of paramount importance for cloud service provision, and load balancing is fully considered when cloud users select an edge cloud service provider (ESP) as it has a direct relation with the quality of cloud service. In multi-cloud systems, a cloud service broker (CSB) reserves cloud resources from multiple CSPs to provision cloud services to users. While existing work has put a lot of attention on IoT applications offloading to the MEC, many of them only considered one edge cloud scenario, ignoring the multi-MEC scenario. In this article, we investigate service pricing and selection for IoT applications offloading in a multi-MEC system with multiple ESPs. Specifically, we take load balancing into account. The studied problem is formulated as a Stackelberg game, where CSB first sets service price and load balancing strategies for the cloud services trying to get its maximized revenue. Then, IoT users make their decisions on which ESP they select service. By applying the backward induction approach, the optimal solutions are derived. The proposed scheme is verified through simulation results.

INDEX TERMS Internet of Things, pricing, load balancing, cloud service broker, edge cloud service provider.

I. INTRODUCTION

In recent years, the Internet of Things (IoT) has received a significant amount of attention in the academic and industrial fields [1]. With the technologies of IoT rapidly growing, various kinds of IoT devices, such as smart phones and vehicles, have been connected by the Internet [2], [3]. The explosive growth of applications generated from the IoT devices has stringent demands of computation resources and real-time processing [4], [5]. However, the IoT devices generally have

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

not enough computing resources, and their battery lives are short. Offloading the IoT applications from the IoT devices to be processed in the servers of remote public cloud is considered as a solution to address the above issue. However, the IoT devices are far away from the public clouds; long transmission time will cost [6]. To overcome this shortcoming, mobile edge computing (MEC) is proposed to solve the challenging problems of limited computing resources and short battery lives of IoT devices [7]. Compared with the public cloud, the edge cloud is near to the IoT devices, therefore, the low latency requirements for the IoT applications can be met. In the last few years, with the demand for edge cloud services proliferating, more and more edge cloud service providers (ESPs) are emerging. For example, China Telecom and Huawei are two ESPs in China.

In the beginning, the model of single ESP dominates the cloud market. With the demands of IoT users' application requests on computing resources increasing dramatically, and at the same time, the applications of IoT users become more and more complex, IoT users may consider using services from different ESPs in order to meet their demands. Under this circumstance, multi-cloud systems, have emerged as practical platforms to address this problem [8].

In a multi-cloud system, a cloud service broker (CSB) acts as a mediator between IoT users and cloud service providers (CSPs). The CSB aggregates integrates and coordinates resources from multiple CSPs and to provide services to IoT users so that not only the demands for cloud services are satisfied, but also that there is enough capacity left for future application requests [9]. A study estimated that the global brokerage market of cloud services had reached 10.5 billion US dollars by the year 2018 [10].

In the cloud system, pricing is of critical importance for service provision, as an important economic factor for CSPs. From the viewpoint of the ESP, how to set its service price is a non-trivial issue. If they pricing services too high, potential IoT users might not select the service from this provider. When the prices are set too low, ESPs might not obtain enough revenue. Many existing works have focused on designing better pricing schemes to maximize the profits or revenue of CSPs in public clouds, such as [11]. From the perspective of CSPs, besides prices, guaranteeing Quality of Service (QoS) measured as the average queueing response time to IoT users is also quite important, as the applications of IoT users need realtime processing [12].

Response time is recognized as an important metric for evaluating the performance of the online cloud services, which is the time between the sending application requests and receiving a response [13]. Besides response time, QoS also includes reliability, security and so on [14]. Response time is used as the main QoS metric, where the assumption is widely made, such as [15]. Besides response time, load balancing is a crucial factor for cloud computing systems. Cloud-based applications have more requirements on load balancing compared to the traditional enterprise applications, and load balancing is an important factor which is fully

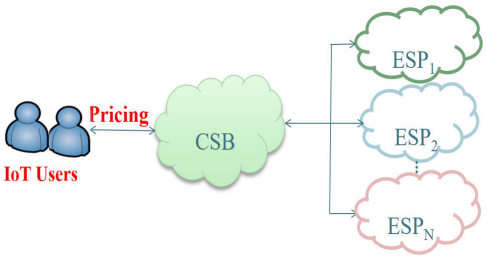

FIGURE 1. The model of a multi-MEC system.

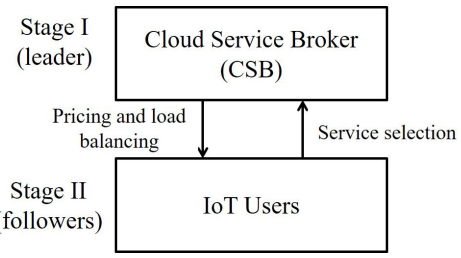

FIGURE 2. The interaction between CSB and IoT users.

considered by users when they choose a CSP [16]. From the perspective of CSPs, load balancing has a direct impact on their revenue as it has a close relation with the quality of service.

Motivated by the concept of multi-cloud systems, this study explores service pricing and selection in a multi-MEC system by taking the different valuations of IoT users' applications on services into account, as shown in Fig.1. The problem that this study tries to address is illustrated in Fig.2. For the CSB, its main objective is to maximize revenue by making optimal pricing and load balancing decisions. On the IoT users' side, based on pricing and load balancing information, they will select services from these ESPs for processing their application requests.

The contributions of this article are summarized as follows.

- Unlike the previous studies that ignored IoT users' different types and load the balancing strategies, this study investigates and analyzes the problem of service pricing and selection for IoT applications offloading in a multi-MEC system taking both pricing and load balancing factors into account.
- Given the prices of cloud services and load balancing strategies of the ESPs in the multi-MEC system, we analyze IoT users' service selection policies based on which we get the effective arrival rate of application request at each ESP. Then, the CSB's revenue maximization problem is formulated. By carefully analyzing the objective function of the problem, we first express the load balancing index as the function of the effective arrival rate of application requests and then propose a dual decomposition method to solve it.
- The proposed method is compared with the proportional scheme (PS) and the resource allocation method of some recent works that did not consider load balancing. The simulation results demonstrate that the proposed method (PM)

can achieve a lower mean response time. We also analyze the effect of reward value on users' arrival rate and the revenue of CSB.

The rest of the paper is structured as follows. Related works are reviewed and discussed in section II. System models are introduced in the section III. The application arrival rate of IoT users' decisions and the CSB's revenue maximization problem are analyzed in the section IV. Simulations are presented in the section V. Section VI presents the conclusions.

II. RELATED WORK

Applications offloading in mobile cloud computing (MCC) and mobile edge computing (MEC) have been extensively in the existing works. In [17], non-orthogonal multiple access (NOMA) is considered as an enabling technology in the future generation communication systems, the authors studied the effect of NOMA on the application offloading in MEC. In [18], the authors developed a unified framework that can minimize the overall outage probability when offloading applications to the public cloud. In [19], Li and Cai proposed an incentive mechanism based framework for the design of collaborative application offloading in MEC. In [20], Pham *et al.* studied the problem of application offloading in the MEC system based on NOMA by using the coalition game. A low-complexity algorithm is developed to get the solution. In [21], Wang *et al.* proposed a multi-antenna NOMA MEC system with multiple users having computation-intensive applications to be offloaded. Minimizing energy consumption from all users is their goal. In [22], Bonadio *et al.* studied the performance of analysis of an SaaS MEC system, which is modelled as a queuing system Markov with multiple servers. However, the authors only considered a single cloud system.

In the cloud systems, pricing provides a useful tool for service provision. Many efforts have been devoted to investigating service price in MEC and MCC without considering load balancing. For example, in [23], the authors proposed a two-stage game-theoretic framework to capture cloud users' demand preferences for cloud capacity and pricing. In [24], Chen *et al.* studied pricing and resource allocation in MEC. They proposed a Stackelberg game-based approach to solving the problems of revenue maximization for MEC and utility maximization for users. In [25], Wu *et al.* investigated how to model the utility of cloud users given the pricing strategy of services of CSPs. In [26], the authors studied task offloading in vehicular fog computing. Two algorithms based on matching-learning and pricing are proposed to minimize the total network delay.

Load balancing is directly related to service quality (e.g., response time) for the cloud users, and it is seriously considered when they select a CSP [16]. Load balancing has been extensively studied in a great number of works with different goals. Nonetheless, many of the existing works ignored pricing factor, which is an intrinsic economic characteristic of cloud services. An energy-aware model is introduced by Paya and Marinescu in [27], which is used for load balancing and application scaling. In [28], Abdelltif *et al.* proposed an SDN-based load balancing service for cloud servers. Their objective was to maximize the utilization of cloud resources and at the same time minimize the response time that cloud users experience. However, they mainly studied from the perspective of cloud users without considering the benefits of CSPs. In [29], Huang *et al.* studied load balancing for caching fairness in edge computing environments by proposing two caching algorithms.

Resource management and service provision in multi-cloud and federated cloud systems have been attracted great attention in recent years. In [30], the authors proposed two algorithms by respectively applying the genetic algorithms and evolutionary game theory for forming the cloud federation. A cooperative cloud market is proposed in [31], where a cloud market broker decides which CSP's resources are employed to be allocated to minimize the cost users. In [32], Li *et al.* proposed a multiple cloud intermediary framework for providing cloud services from different CSPs to users for streaming big data processing. Lin *et al.* proposed a brokerage-based framework in the cloud computing systems, where the CSB takes the responsibility of recommending the best available services to users [33]. In [34], Mei *et al.* studied the problem of profit maximization for a cloud broker. The cloud broker buys reserved VM instances from a multiple of cloud providers and delivers services to the users. However, these previous studies neglected user type heterogeneity and response time, which are not practical as different types of applications may have different preferences for cloud services and different requirements for processing deadlines. In [35], the authors presented a novel framework to secure the access to IoT services in a multi-cloud system without considering user heterogeneity, pricing factor and load balancing strategy. In [36], Zhou *et al.* studied the problem of edge computing service provisioning for energy-efficient workload offloading in vehicular networks, and proposed a consensus ADMM approach to solve it. However, these works did not consider the pricing factor.

From the analysis of the works above, it can be found that many of the previous works on IoT applications offloading did not jointly study pricing and load balancing. Furthermore, these previous works neglected IoT users' heterogeneity, as different IoT users types may show different preferences for cloud services [11]. This study compares with several closely related work on service provision in multi-cloud and federated cloud systems, which is summarized in Table 1. "✓" means that this factor is considered and "×" means that this factor is ignored. It can be observed from Table 1 that the main related studies only considered part of the factors, whereas, our study fully consider these factors.

III. SYSTEM MODELS

This section introduces the system models consisting of IoT users and CSB. Consider a multi-MEC system with a CSB who reserves cloud resources from multiple ESPs and provisions cloud services to IoT users. Each IoT user will make a

TABLE 1. A comparison with main related work.

Related Work	User Utility	User Heterogeneity	Price Factor	Service Quality	Load Balancing
[37]	×	×	×	✓	×
[38]	×	×	✓	×	×
[33]	×	×	×	×	×
[34]	×	×	×	×	×
[35]	×	×	✓	✓	×
[32]	✓	×	✓	✓	×
[22]	×	×	✓	×	×
This Work	✓	✓	✓	✓	✓

determination on which ESP's service it should select from to process its application requests.

A. CSB's MODEL

We suppose that the number of ESPs that a CSB integrates and coordinates resources from is N, each of which is modelled by one M/M/1 queue, the processing capacity of whom is represented by its service rate μ_i (in application requests per second), $i \in \{1, 2, \ldots, N\}$. Without loss of generality, it is also assumed that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_N \geq 0$. For each ESP$_i$, CSB sets the price per application request as p_i, and broadcasts load balancing indicator s_i to the IoT users. Based on this information, IoT users will decide whether to subscribe to cloud services from one of the ESPs or not.

B. THE MODEL OF IoT USERS

We made such an assumption that IoT users' application requests are generated with the arrival rate λ (measured by the application requests per second) according to Poisson process [36], and these application requests are processed based on the first-come-first-served (FCFS) queueing discipline. The recent studies on the analysis of cloud data centers show that the arrival rate of users' application requests follows Poisson distribution [15], [39]. We also assume that each IoT user generates a distinct application request upon arrival. Therefore, the number of application requests reflects the number of IoT users. IoT user and application request are used interchangeably throughout the paper.

Remark: It is important to remark that an M/GI/1 PS system can also be adopted or an M/M/1 FCFS system. However, the problem can be simplify analyzed by considering an M/M/1 FCFS system. The detailed reason can be referred to the work of [40].

Each IoT user is assumed to be associated with a specific application request by the parameter α_k having uniform distribution in the range [0, 1], and its probability distribution function (PDF) and cumulative distribution function (CDF) are denoted as $f(\cdot)$ and $F(\cdot)$, respectively [11]. The value of α_k represents this IoT user's preference for cloud services. Based on the above assumptions, given the service price p_i and load balancing indicator s_i of ESP$_i$, the utility that this IoT user α_k gets is

$$U_{k,i} = \alpha_k r - c d_i(\lambda) - p_i, i = 1, 2, \ldots, N \quad (1)$$

where r denotes the reward that this IoT user obtains from using cloud service, and it is proportional to the application request size of this IoT user [23], [41], c is the delay cost which reflects the urgency of users' application requests [15], [23], and $d_i(\lambda_i)=1/(\mu_i - \lambda_i)$ is the average response time (including queueing time plus processing time) incurred by the arrival rate λ_i. As μ_i and λ_i are both measured by application requests per second; therefore, the average response time is measured by seconds per application request. It is obvious that serving an application request with larger size means that this IoT user can gets a higher reward.

For ease of presentation, the notations of this article are summarised in Table 2.

TABLE 2. Notations summary.

Notation	Description
N	the number of ESPs in the multi-cloud system
r	the reward that each IoT user obtains
p_i	the service price of ESP$_i$, for $i = 1, 2, ..., N$
d_i	average queueing delay of ESP$_i$
λ	the number of IoT users' application requests
μ_i	the service rate of ESP$_i$
c	the delay cost per unit time
λ_i	the number of IoT users that chooses ESP$_i$
s_i	the load balancing strategy of ESP$_i$
α_k	the type of IoT user k
$U_{k,i}$	the utility that the IoT user type k gets from ESP$_i$, for $i = 1, 2, ..., N$

IV. PROBLEM FORMULATION AND SOLUTION METHOD

In this section, the interaction between IoT users and CSB are investigated and analyzed by applying a two-stage Stackelberg game [42], the illustration of which is shown in Fig.2. In this game, the CSB sets service prices and load balancing strategies to have the maximized revenue in the first stage, and IoT users make determinations on the number of their generated application requests in the second stage. Based on the method of backward induction, this game can be solved easily. In the first stage, based on service prices and load balancing indicators that set by the CSB, we study how IoT users make their determinations on which ESP's cloud service should be selected and the number of application requests they generate. In the second stage, we study how CSB maximizes its revenue according to IoT users' decisions in the first stage.

A. STAGE II: IoT USERS' STRATEGIES

Recall that the arrival rate of IoT users' application requests is rate λ, which is assumed to follow poisson process, and these application request are executed based on the discipline of FCFS queueing.

For the type α_k IoT user, after generating application requests, a decision will be made as to which ESP's service it selects from. It joins the ESP only if the utility it gets is positive, which means that

$$U_{k,i} = \alpha_k r - c d_i(\lambda_i) - p_i > 0 \quad (2)$$

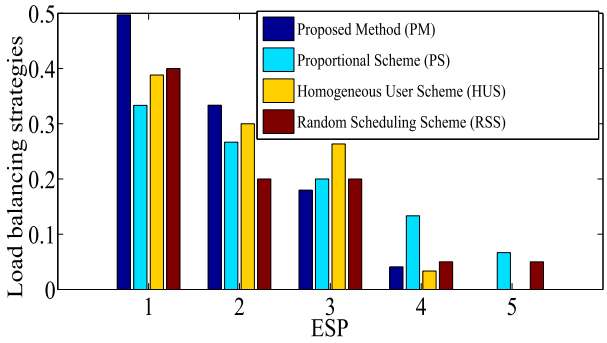

FIGURE 3. Comparison with benchmark methods in terms of the mean response time.

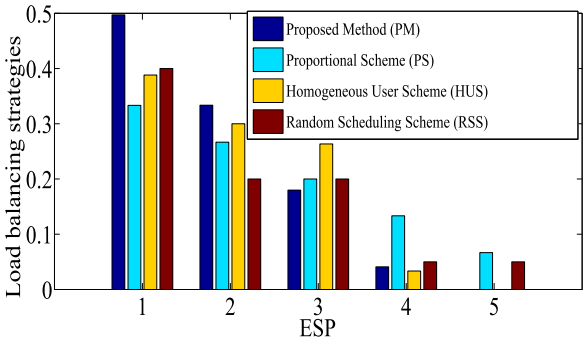

FIGURE 4. Comparison with benchmark methods in terms of the load balancing strategies.

from which we get $\alpha_k > \alpha_k^*$ where

$$\alpha_k^* = \frac{cd_i(\lambda_i) + p_i}{r} \tag{3}$$

Therefore, the effective application request arrival rate of IoT user α_k at ESP_i is

$$\begin{aligned}\lambda_i &= s_i\lambda[\int_{\alpha_k^*}^{1} f(\alpha)d(\alpha)] \\ &= s_i\lambda[1 - F(\alpha_k^*)] \\ &= s_i\lambda[1 - \frac{cd_i(\lambda_i(s_i)) + p_i}{r}]\end{aligned} \tag{4}$$

where s_i denotes the fraction of the number of application requests dispatched to ESP_i, and denotes the load balancing indicator of this ESP.

B. STAGE I: CSB's REVENUE MAXIMIZATION

After the effective arrival rate λ_i is got, the problem of revenue optimization for CSB is

Problem 1:

$$\begin{aligned}&\max \sum_{i=1}^{N} p_i\lambda_i \\ &s.t.\ \lambda_i = s_i\lambda[1 - \frac{cd_i(\lambda_i) + p_i}{r}] \\ &\sum_{i=1}^{N} s_i = 1 \\ &0 \le s_i \le 1 \\ &p_i \ge 0 \\ &variables\{p_i, s_i\}\end{aligned} \tag{5}$$

where the first constraint is the effective application arrival rate of IoT user α_k at ESP_i, and the second one is the constraint of load balancing.

By solving **Problem1**, the following results can be obtained, the proof of which are shown in **Appendix**.

Proposition 1. The unique optimal pricing and load balancing strategies from solving the revenue maximization problem are

$$s_i^* = \frac{\lambda_i(v^*)}{\lambda_t(v^*)}, i = 1, 2, \ldots, N \tag{6}$$

$$p_i^* = r - \frac{r\lambda_i(s_i^*)}{\lambda s_i^*} - cd_i(\lambda_i(s_i^*)), \quad i = 1, 2, \ldots, N \tag{7}$$

V. SIMULATION RESULTS

Consider a multi-MEC system with a CSB who reserves cloud resource from $N = 5$ ESPs and provisions edge cloud services to a number of IoT users. We present simulation results to evaluate and compare our proposed method (PM) with some main recent work on cloud service provision.

We compare the following benchmark schemes with our algorithm:

Proportional Scheme (PS): PS dispatches application requests of IoT users to ESPs according to their processing rates [43]:

$$\lambda_i = \lambda \frac{\mu_i}{\sum_{i=1}^{N} \mu_i} \tag{8}$$

Homogeneous User Scheme (HUS): HUS assumes that all the IoT users have a homogeneous preference for cloud services.

Random Scheduling Scheme (RSS): RRS dispatches application requests of IoT users to ESPs randomly. We assume that the number of application requests dispatched to the ESPs according to the following load balancing strategies: $s_1 = 2/5, s_2 = 1/5, s_3 = 1/5, s_4 = 1/20, s_5 = 1/20$. The IoT users are assumed to be homogeneous under PS. We analyze the effect of the reward value, the effect of arrival rate, and the effect of system size on CSB's revenue and IoT user's total utility.

A. PARAMETER VALUES SETTING

For a multi-MEC system, the default values of the parameters are: The values of $r = 2$, $c = 1$, $N = 5$, μ_i $(i = 1, 2, \ldots, 5)$ are set respectively as 10, 8, 6, 4, 2, and $\lambda = 20$.

The Discussion of the setting for parameter values. We let μ_i denote the service rate for the ESP_i reflecting the application request this provider's processing capacity.

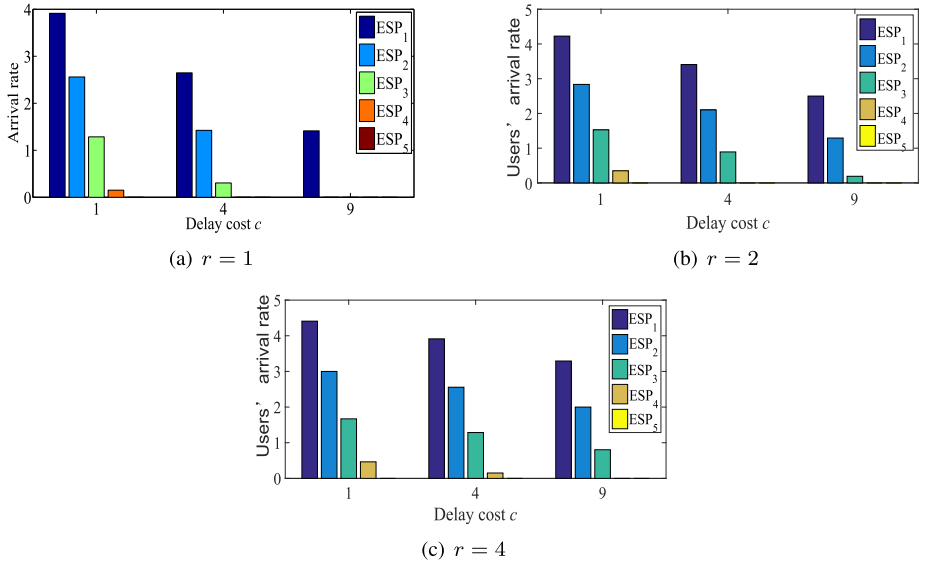

FIGURE 5. The effect of reward value r on IoT users' arrival rate.

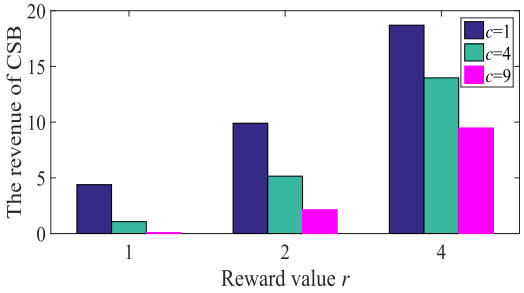

FIGURE 6. The effect of reward value r on CSB's revenue.

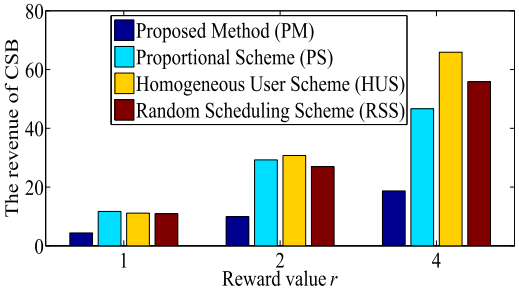

FIGURE 7. Comparison with benchmark methods in terms of the revenues under different reward values.

The value of λ denotes the number of application requests. r is the reward of this IoT user obtained from using edge cloud services, which is directly related with the data size of the application request. It is obvious that processing an application request with larger size means that this IoT user can get more reward. Similar to the work of [44], we let the value of r represent the size of application request of the cloud user. The varying values of r reflect the different sizes of application requests. The values selected for the parameters are only for the purpose of demonstration. Similar results can also be obtained by choosing other values of parameters.

B. COMPARISON OF MEAN RESPONSE TIME

We first compare PM with the benchmark methods in terms of mean response time that IoT users experience in each ESP and the revenue that each ESP gets. For the HUS, we set the values of all the IoT users' types as $\alpha_k = 1$. Fig.3 shows the comparison of the PM with PS, HUS and RSS in terms of the mean response time experienced in each ESP. From Fig.3, we can see that PM achieves much lower response time compared with the benchmark methods, which means that the PM can provide cloud services with better QoS. Fig.3 also suggests that these ESPs with more cloud resources can provision services with better QoS than those with less cloud capacities in the multi-MEC system.

C. COMPARING LOAD BALANCING STRATEGIES

We next compare load balancing strategies set by CSB for different ESPs under our proposed method PM and PS. From Fig.4, we observe that the PM assigns more service requests to the ESPs with more cloud resources than these ESPs with fewer ones. One main reason is that PM considers IoT users' different types, which are ignored in PS.

D. EFFECT OF REWARD VALUE

We analyze how reward value r impacts IoT users' arrival rate of application requests and CSB's revenue. Figs.5(a), (b) and (c) depict the impact of reward value r on IoT users' arrival rate at each ESP with different delay costs in the multi-MEC system. It is easily seen from the three figures that IoT users with larger sizes of application requests are prone to use cloud services under the condition that delay costs are the same, and IoT users are more likely to choose

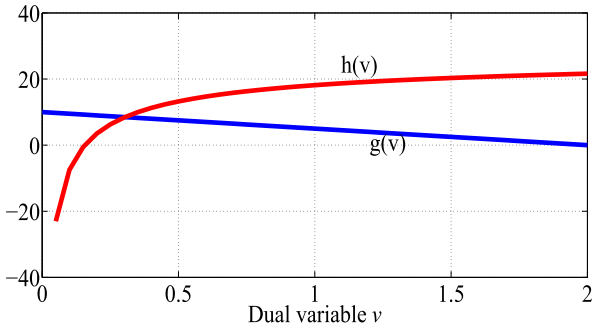

FIGURE 8. Numerical analysis of dual variable v.

these ESPs who provide low delay costs. These figures also indicate that these ESPs with abundant cloud resources can process more application requests from IoT users than those with less cloud capacities as delay costs increase. No IoT user chooses ESP_5 due to its not enough cloud resources.

The impact of reward value r on CSB's revenue is shown in Fig.6. It is observed in this figure that the revenue of CSB increases with reward value increasing, which indicates that CSB can gain more revenue if the provided edge cloud services can process requests with higher sizes. Fig.6 also shows that higher delay costs will lead to the loss of revenue.

Fig.7 shows the comparison of our proposed scheme with benchmark method under different reward values. This figure suggests that CSB earns the least revenue under the proposed schemes comparing with benchmark methods. This is due to the reason that the proposed scheme considers IoT users' different tastes for cloud services, which is more practical; in contrast, the benchmark methods assume that all the IoT users have the same valuation on cloud services.

VI. CONCLUSION AND FUTURE WORKS

We have presented a joint study of the problem of service pricing and selection for IoT applications offloading in a multi-MEC system. The problem is investigated and analyzed as a Stackelberg game, under which the CSB first sets service prices and load balancing strategies for the purpose of maximizing revenue. Then, IoT users makes their application requests arrival rate decisions based on the information provided by CSB. This game is solved by applying the technique of backward induction, in which the CSB determines service prices and load balancing strategies in Stage II and IoT users decide which ESP's service they choose in Stage I. The simulation results demonstrate that in a comparison with the benchmark methods, the proposed method (PM) achieves much lower mean response time and provides a more effective load balancing strategy.

Several research aspects can still be further explored as future works. For example, the ESPs in the multi-MEC system may compete or cooperate to provision edge cloud services. Therefore, considering the relationship of these ESPs is an interesting future work. Our work can also be extended to analyze IoT users' different sensitivity to delay into account.

APPENDIX

PROOF 1

The objective function of the revenue maximization problem is not convex, and it is difficult to be solved.

According to the first constraint, we have

$$p_i(s_i, \lambda_i) = r - \frac{r\lambda_i(s_i)}{\lambda s_i} - cd_i(\lambda_i(s_i)) \tag{9}$$

Therefore, we can eliminate the first constraint of **Problem1** by substituting Eq.(9) into the objective function, and the original problem is equivalently written as,

Problem 2:

$$\begin{aligned} \max \ & \sum_{i=1}^{N}[r\lambda_i - \frac{r(\lambda_i)^2}{\lambda} - \lambda_i cd_i(\lambda_i)] \\ s.t. \ & \sum_{i=1}^{N} s_i = 1 \\ & 0 \le s_i \le 1 \\ & \lambda_i \ge 0 \\ & variables\{s_i, \lambda_i\} \end{aligned} \tag{10}$$

Before introducing solution method, we first give a lemma which is proved in [45], based on which our solution method is proposed.

Lemma 1. We always have $\underset{x,y}{supf}(x, y)=\underset{x}{sup\tilde{f}}(x)$,

where $\tilde{f}(x)=sup_y f(x, y)$.

Lemma 1: means that we could always first minimize a function through minimizing over some of the variables, and minimizing the remaining ones later [45].

According to **Lemma 1**, we can solve **Problem2** by maximizing over s_i and λ_i sequentially. For any given λ_i, the above problem is transformed into

Problem 3:

$$\begin{aligned} \max \ & \sum_{i=1}^{N}[-\frac{r(\lambda_i)^2}{\lambda s_i}] \\ s.t. \ & 0 \le s_i \le 1 \\ & \sum_{i=1}^{N} s_i = 1 \\ & variable\{s_i\} \end{aligned} \tag{11}$$

It is obvious that **Problem3** is a convex problem, from KKT conditions [45], the solution is obtained as follows

$$s_i = \frac{\lambda_i}{\sum_{i=1}^{N} \lambda_i}, \quad i = 1, 2, \ldots, N \tag{12}$$

By substituting Eq.(12) back into Eq.(10) and denoting $\sum_{n=1}^{N} \lambda_i = \lambda_t$, we have an equivalent problem

Problem 4:

$$\begin{aligned} &\max[r\lambda_t - \frac{r(\lambda_t)^2}{\lambda} - \sum_{n=1}^{N} \lambda_i c d_i(\lambda_i)] \\ &s.t. \sum_{i=1}^{N} \lambda_i = \lambda_t \\ &\lambda_i \geq 0 \\ &\lambda_t \geq 0 \\ &variables\{\lambda_t, \lambda_i\} \end{aligned} \tag{13}$$

The objective function of **Problem4** is convex with respect to λ_i, and the Lagrangian function is

$$L(\lambda_t, \lambda_i, v) = r\lambda_t - \frac{r(\lambda_t)^2}{\lambda} - \sum_{n=1}^{N} \lambda_i c d_i(\lambda_i) + v(\sum_{n=1}^{N} \lambda_i - \lambda_t) \tag{14}$$

where v denotes the dual variable associating with the first constraint of **Problem4**. Denoting

$$L_1(\lambda_t, v) = r\lambda_t - \frac{r(\lambda_t)^2}{\lambda} - v\lambda_t \tag{15}$$

$$L_2(\lambda_i, v) = -\lambda_i c d_i(\lambda_i) + v\lambda_i \tag{16}$$

It is obvious that Eq.(15) is a concave function of λ_t for a given v, and Eq.(16) is a concave function of λ_i for a given v. From the first-order condition [45], we have

$$\frac{\partial L_1}{\partial \lambda_t} = r - \frac{2r\lambda_t}{\lambda} - v = 0 \tag{17}$$

$$\frac{\partial L_2}{\partial \lambda_i} = \frac{-c\mu_i}{(\mu_i - \lambda_i)^2} + v = 0 \tag{18}$$

Therefore, the optimal solutions of Eqs.(15) and (16) for a given v are respectively expressed as

$$\lambda_t(v) = [g(v)]^+ \tag{19}$$

$$\lambda_i(v) = [h(v)]^+ \tag{20}$$

where $[x]^+$ is $\max\{0, x\}$, $g(v)$ and $h(v)$ are given as

$$g(v) = \frac{\lambda(r - v)}{2r} \tag{21}$$

$$h(v) = \mu_i - \sqrt{\frac{c\mu_i}{v}} \tag{22}$$

It is evident that for $v \in (c/\mu_i, r)$, $h(v)$ is an increasing and positive function, and $g(v)$ is a decreasing and positive function. Therefore, we can get the optimal solutions of $\lambda_t(v)$ and $\lambda_i(v)$ by calculating the optimal value of dual variable v^*. By substituting Eqs. (21) and (22) into the first constraint of **Problem4**, we have

$$\sum_{i=1}^{N} (\mu_i - \sqrt{\frac{c\mu_i}{v}})^+ = \frac{\lambda(r - v)}{2r} \tag{23}$$

Eq.(23) is difficult to be solved, but we can try to solve it by numerical analysis. As illustrated in Fig.3, the two lines intersect at a unique point where the optimal value of v is obtained. The values of $r = 2$, $c = 1$, $N = 5$, μ_i $(i = 1, 2, \ldots, 5)$ are set respectively as 10, 8, 6, 4, 2, and $\lambda = 20$.

After getting the unique v^*, we can obtain the unique solutions $\lambda_i(v^*)$ and $\lambda_t(v^*)$ from Eqs.(14) and (15), respectively, which are also global optimal solutions of **Problem4**, as v^*, $\lambda_i(v^*)$ and $\lambda_t(v^*)$ satisfy the sufficient and necessary KKT conditions [45]. By substituting these values into Eqs.(6) and (9), we get s_i^* and p_i^*, respectively. Therefore, **Proposition 1** is proved.

REFERENCES

[1] S. Hu and G. Li, "Dynamic request scheduling optimization in mobile edge computing for IoT applications," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1426–1437, Feb. 2020.

[2] Z. Zhou, H. Liao, B. Gu, S. Mumtaz, and J. Rodriguez, "Resource sharing and task offloading in IoT fog computing: A contract-learning approach," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 3, pp. 227–240, Jun. 2020.

[3] W. Zhang, X. Li, L. Zhao, and X. Yang, "Competition of duopoly MVNOs for IoT applications through wireless network virtualization," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–11, May 2020.

[4] S. Hu and G. Li, "Dynamic request scheduling optimization in mobile edge computing for IoT applications," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1426–1437, Feb. 2020.

[5] S. Chen, L. Zhang, Y. Tang, C. Shen, R. Kumar, K. Yu, U. Tariq, and A. K. Bashir, "Indoor temperature monitoring using wireless sensor networks: A SMAC application in smart cities," *Sustain. Cities Soc.*, vol. 61, Oct. 2020, Art. no. 102333.

[6] Y. Wang, C. Xu, Z. Zhou, H. Pervaiz, and S. Mumtaz, "Contract-based resource allocation for low-latency vehicular fog computing," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Bologna, Italy, Sep. 2018, pp. 812–816.

[7] Y. Wu, L. P. Qian, K. Ni, C. Zhang, and X. Shen, "Delay-minimization nonorthogonal multiple access enabled multi-user mobile edge computation offloading," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 392–407, Jun. 2019.

[8] S. Sugawara, "Survey of cloud-based content sharing research: Taxonomy of system models and case examples," *IEICE Trans. Commmu.*, vol. E100.B, no. 4, pp. 484–499, Apr. 2017.

[9] L. Mashayekhy, M. M. Nejad, and D. Grosu, "Cloud federations in the sky: Formation game and mechanism," *IEEE Trans. Cloud Comput.*, vol. 3, no. 1, pp. 14–27, Jan./Mar. 2015.

[10] R. Zhang, K. Wu, M. Li, and J. Wang, "Online resource scheduling under concave pricing for cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 4, pp. 1131–1145, Apr. 2016.

[11] L. Tang and H. Chen, "Joint pricing and capacity planning in the IaaS cloud market," *IEEE Trans. Cloud Comput.*, vol. 5, no. 1, pp. 158–171, Jan./Mar. 2017.

[12] M. Camelo, M. Claeys, and S. Latré, "Parallel reinforcement learning with minimal communication overhead for IoT environments," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1387–1400, Feb. 2020.

[13] J. Zhu, Z. Zheng, and M. R. Lyu, "DR2: Dynamic request routing for tolerating latency variability in online cloud applications," in *Proc. IEEE 6th Int. Conf. Cloud Comput. (CLOUD)*, Santa Clara, CA, USA, Jun. 2013, pp. 589–596.

[14] Y. Chen, L. Wang, X. Chen, R. Ranjan, A. Zomaya, Y. Zhou, and S. Hu, "Stochastic workload scheduling for uncoordinated datacenter clouds with multiple QoS constraints," *IEEE Trans. Cloud Comput.*, early access, Jun. 29, 2016, doi: 10.1109/TCC.2016.2586048.

[15] C. T. Do, N. H. Tran, E.-N. Huh, C. S. Hong, D. Niyato, and Z. Han, "Dynamics of service selection and provider pricing game in heterogeneous cloud market," *J. Netw. Comput. Appl.*, vol. 69, pp. 152–165, Jul. 2016.

[16] K. Li, "Optimal load distribution for multiple heterogeneous blade servers in a cloud computing environment," *J. Grid Comput.*, vol. 11, no. 1, pp. 27–46, Mar. 2013.

[17] Z. Ding, P. Fan, and H. V. Poor, "Impact of non-orthogonal multiple access on the offloading of mobile edge computing," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 375–390, Jan. 2019.

[18] D. Han, W. Chen, B. Bai, and Y. Fang, "Offloading optimization and bottleneck analysis for mobile cloud computing," *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 6153–6167, Sep. 2019.

[19] G. Li and J. Cai, "An online incentive mechanism for collaborative task offloading in mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 624–636, Jan. 2020.

[20] Q.-V. Pham, H. T. Nguyen, Z. Han, and W.-J. Hwang, "Coalitional games for computation offloading in NOMA-enabled multi-access edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1982–1993, Feb. 2020.

[21] F. Wang, J. Xu, and Z. Ding, "Multi-antenna NOMA for computation offloading in multiuser mobile edge computing systems," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2450–2463, Mar. 2019.

[22] A. Bonadio, F. Chiti, and R. Fantacci, "Performance analysis of an edge computing SaaS system for mobile users," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 2450–2463, Mar. 2020.

[23] M. Taghavi, J. Bentahar, and H. Otrok, "Two-stage game theoretical framework for IaaS market share dynamics," *Future Gener. Comput. Syst.*, vol. 102, pp. 173–189, Jan. 2020.

[24] Y. Chen, Z. Li, B. Yang, K. Nai, and K. Li, "A Stackelberg game approach to multiple resources allocation and pricing in mobile edge computing," *Future Gener. Comput. Syst.*, vol. 108, pp. 273–287, Jul. 2020.

[25] C. Wu, R. Buyya, and K. Ramamohanarao, "Modeling cloud business customers' utility functions," *Future Gener. Comput. Syst.*, vol. 105, pp. 737–753, Apr. 2020.

[26] H. Liao, Z. Zhou, X. Zhao, B. Ai, and S. Mumtaz, "Task offloading for vehicular fog computing under information uncertainty: A matching-learning approach," in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Tangier, Morocco, Jun. 2019, pp. 2001–2006.

[27] A. Paya and D. C. Marinescu, "Energy-aware load balancing and application scaling for the cloud ecosystem," *IEEE Trans. Cloud Comput.*, vol. 5, no. 1, pp. 15–27, Jan./Mar. 2017.

[28] A. A. Abdelltif, E. Ahmed, A. T. Fong, A. Gani, and M. Imran, "SDN-based load balancing service for cloud servers," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 106–111, Aug. 2018.

[29] Y. Huang, X. Song, F. Ye, Y. Yang, and X. Li, "Fair and efficient caching algorithms and strategies for peer data sharing in pervasive edge computing environments," *IEEE Trans. Mobile Comput.*, vol. 19, no. 4, pp. 852–864, Apr. 2020.

[30] A. Hammoud, A. Mourad, H. Otrok, O. A. Wahab, and H. Harmanani, "Cloud federation formation using genetic and evolutionary game theoretical models," *Future Gener. Comput. Syst.*, vol. 104, pp. 92–104, Mar. 2020.

[31] K. H. K. Reddy, G. Mudali, and D. Sinha Roy, "A novel coordinated resource provisioning approach for cooperative cloud market," *J. Cloud Comput.*, vol. 6, no. 1, pp. 1–17, Dec. 2017.

[32] H. Li, M. Dong, K. Ota, and M. Guo, "Pricing and repurchasing for big data processing in multi-clouds," *IEEE Trans. Emerg. Topics Comput.*, vol. 4, no. 2, pp. 266–277, Apr. 2016.

[33] D. Lin, A. C. Squicciarini, V. N. Dondapati, and S. Sundareswaran, "A cloud brokerage architecture for efficient cloud service selection," *IEEE Trans. Services Comput.*, vol. 12, no. 1, pp. 144–157, Jan. 2019.

[34] J. Mei, K. Li, Z. Tong, Q. Li, and K. Li, "Profit maximization for cloud brokers in cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 1, pp. 190–203, Jan. 2019.

[35] M. Kazim, L. Liu, and S. Y. Zhu, "A framework for orchestrating secure and dynamic access of IoT services in multi-cloud environments," *IEEE Access*, vol. 6, pp. 58619–58633, 2018.

[36] Z. Zhou, J. Feng, Z. Chang, and X. Shen, "Energy-efficient edge computing service provisioning for vehicular networks: A consensus ADMM approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5087–5099, May 2019.

[37] J. L. Lucas-Simarro, R. Moreno-Vozmediano, R. S. Montero, and I. M. Llorente, "Scheduling strategies for optimal service deployment across multiple clouds," *Future Gener. Comput. Syst.*, vol. 29, no. 6, pp. 1431–1441, Aug. 2013.

[38] W. Wang, D. Niu, B. Liang, and B. Li, "Dynamic cloud instance acquisition via IaaS cloud brokerage," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 6, pp. 1580–1593, Jun. 2015.

[39] M. Liu, W. Dou, S. Yu, and Z. Zhang, "A decentralized cloud firewall framework with resources provisioning cost optimization," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 3, pp. 621–631, Mar. 2015.

[40] A. Wierman, L. L. H. Andrew, and A. Tang, "Power-aware speed scaling in processor sharing systems: Optimality and robustness," *Perform. Eval.*, vol. 69, no. 12, pp. 601–622, Dec. 2012.

[41] Y. C. Lee, C. Wang, A. Y. Zomaya, and B. B. Zhou, "Profit-driven scheduling for cloud services with data access awareness," *J. Parallel Distrib. Comput.*, vol. 72, no. 4, pp. 591–602, Apr. 2012.

[42] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA, USA: MIT Press, 1991.

[43] D. Grosu and A. T. Chronopoulos, "Noncooperative load balancing in distributed systems," *J. Parallel Distrib. Comput.*, vol. 65, no. 9, pp. 1022–1034, Sep. 2005.

[44] J. Chen, C. Wang, B. B. Zhou, L. Sun, Y. C. Lee, and A. Y. Zomaya, "Tradeoffs between profit and customer satisfaction for service provisioning in the cloud," in *Proc. 20th Int. Symp. High Perform. Distrib. Comput. HPDC*, San Jose, CA, USA, Jun. 2011, pp. 229–238.

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

WANLI ZHANG received the B.S. degree from China Agricultural University and the M.S. degree from Anhui University. He is currently an Associate Professor with the School of Information Engineering, Suzhou University. His main research interests include wireless sensor networks, the Internet of Things, mobile edge computing, cloud computing, and computer networks.

XIANWEI LI received the M.S. degree from Hunan University, Changsha, China, in 2010, and the Ph.D. degree from Waseda University, Tokyo, Japan, in 2019. Since July 2011, he has been with the School of Information Engineering, Suzhou University. His research interests include mobile edge computing, the Internet of Things, machine learning, and big data.

LIANG ZHAO received the Ph.D. degree from the School of Computing, Edinburgh Napier University, in 2011. He is currently an Associate Professor with Shenyang Aerospace University, China. Before that, he worked as an Associate Senior Researcher with Hitachi Research and Development Corporation, China, from 2012 to 2014. His research interests include WMNs, VANETs, MANETs, and intelligent transport services.

XIAOYING YANG received the received the master's degree of computer technology from the Guilin University of Electronic Technology, in 2010. She is currently an Associate Professor with the School of Information Engineering, Suzhou University, Anhui, China. Her research interests include wireless sensor networks, the Internet of Things, and database systems.

TAO LIU received the B.S. degree in computer application technology from the Hefei University of Technology, China, in 2004. She was a Visiting Scholar with the University of Science and Technology of China, in 2012. She is currently a Professor with the School of Computer and Information, Anhui Polytechnic University. Her main research interests include machine learning, computer networks, and information security.

WEI YANG received the bachelor's degree in computer science and technology from Anhui Normal University, in 2004, and the master's degree in computer technology from Anhui University, in 2014. She is currently a Lecturer with the School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, China. Her research interests include big data and data mining.