Pricing Policy and Computational Resource Provisioning for Delay-aware Mobile Edge Computing

Tianchu Zhao, Sheng Zhou, Xueying Guo, Yun Zhao, Zhisheng Niu
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
Email: zhaotc13@mails.tsinghua.edu.cn, sheng.zhou@tsinghua.edu.cn, guo-xy11@mails.tsinghua.edu.cn
zhaoyun12@mails.tsinghua.edu.cn, niuzhs@tsinghua.edu.cn

***Abstract*—Mobile edge computing is a novel technique to offer cloud-based computation offloading services to mobile users with short delay. However, the Cloud Service Providers (CSPs) of the edge cloud are generally different from the CSPs of remote Internet cloud. Considering the competition between the heterogeneous clouds, we study the optimal provisioning of the computational resource in the edge cloud. We firstly analyze the Nash equilibrium prices of the cloud market when the amount of edge computational resource is given. Based on the pricing policy, we further design an algorithm to optimize the edge computational resource capacity so that the profit of the edge cloud is maximized. We also derive the lower and upper bound of the optimal edge computational resource. The numerical results indicate that the profit of the edge cloud is greatly influenced by the price of the remote cloud when users care more about the price than the delay.**

I. Introduction

Mobile devices and applications are more and more widely used in recent years. However, the limited battery capacity and resource are becoming the bottleneck of better Quality of Experience (QoE) [1]. Mobile Cloud Computing (MCC) has been proposed to solve this problem. By offloading computational tasks to cloud servers over wireless and wired networks, mobile devices could greatly save energy and improve performance [2]. But the data transmission delay over the Internet is generally long, which makes it hard to guarantee the delay requirements for real-time services [3][4]. Therefore, Mobile Edge Computing (MEC) is proposed to realize efficient computation offloading with short delay [5]. The cloud that is deployed in the edge of the network avoids the Internet data transmission between users and cloud servers. The edge cloud provides real-time services, and many architectures are proposed accordingly(e.g. cloudlet [6], femtocloud [7], CONCERT [8]).

Both the edge cloud and the remote Internet cloud provide computation offloading services to mobile users, which usually belong to different CSPs. For example, Huawei is an emerging edge cloud providers [5] while Amazon EC2 is a traditional remote cloud [9]. In the MCC market, the heterogeneous clouds provide similar services, and they compete with each other for mobile users. The CSP of the edge cloud should offer qualified services so that it could attract users. Thus, enough computational resource is needed to be deployed in the edge cloud to meet the requirements of users.

The monetary cost to deploy and operate data centers in the edge cloud is another concern of the CSPs. In [10], it is shown that about half of the total cost of the remote Internet cloud is consumed to purchase servers which charges $\$3000$ each. Moreover, about 25% of the total cost is used to power and cool the servers. Compared with the centralized remote cloud, the edge clouds are distributed and densely deployed with smaller scale, which results in higher cost. In [11], it is noted that the cost per MIPS (million instructions per second) to operate the cloud is highly related to the scale. As the cloud with larger scale reduce the cost by achieving higher pooling gain, the distributed deployed edge clouds suffer from large amount of investment. Thus, the provisioning of the edge cloud is a joint consideration of both the demand of users and the cost.

In our previous work [12], we study the relationship between the amount of computational resource in the cloud and the QoE. We find that the QoE is weakly improved by deploying more computational resource if the total resource exceeds a threshold. But the cost is not considered, which is a quite important factor when provisioning computational resource in the edge cloud. In [13], the authors study the pricing policy of multiple CSPs in a MCC market, and they design an algorithm to derive the prices with low complexity. But they focus on prices of homogeneous clouds, which is not practical in the mobile edge computing scenario including both the edge cloud and the remote cloud. In [14], the authors study the pricing policy of CSPs whose computational resource are already given. But they fail to further analyze how the amount of computational resource influence the profits of CSPs.

In this paper, we consider a MCC market in which both the edge cloud and the remote cloud exist (Fig. 1), and study how much computational resource should be deployed in the edge cloud so that its profit is maximized. The profit of the edge cloud is related to many factors which have already been discussed. To optimize this problem, we firstly derive the pricing policy of heterogeneous clouds when the

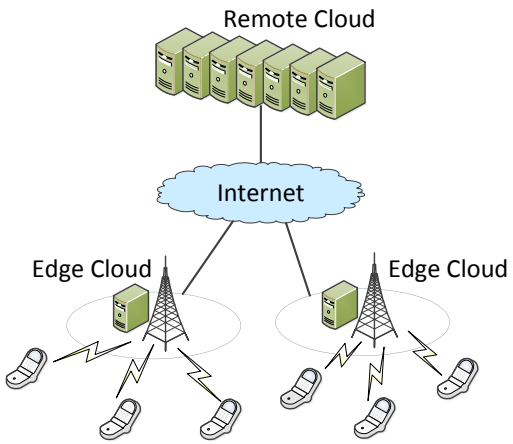


Fig. 1. A typical MCC scenario with heterogeneous clouds. Mobile users can offload tasks to the edge cloud over wireless and backhaul network. They can also offload tasks to the remote cloud over the Internet.

computational resource in the edge cloud is given, which are indeed the Nash equilibrium prices. Based on the close formulation of prices, the problem is then optimized by a proposed algorithm. We also derive the lower and upper bound of the optimal computational resource which maximize the profit of the edge cloud.

The rest of the paper is organized as follows. Section II introduces the system model and derives the problem formulation. In Section III, we study the Nash equilibrium prices of heterogeneous clouds. In Section IV, we optimize the provisioning of computational resource in the edge cloud. In Section V, the numerical results is shown to validate our analysis. The paper is concluded in Section VI.

II. System Model and Problem Formulation

We consider a mobile cloud computing market with the edge cloud and the remote cloud. Among the clouds, a mobile user choose one of them each time to offload the computational task. Necessary data is firstly transmitted to data centers, and the execution will begin afterwards. The edge cloud is deployed near mobile users, and data is transmitted to data centers through wireless channel and backhaul which is a wired link connecting the base station and the edge cloud. Compared with the edge cloud, remote execution needs an additional Internet data transmission. As the edge cloud has limited computational resource, tasks might need to wait in a queue before they are executed. The remote cloud has abundant computational resource, and the execution of each task will immediately begin as long as the data transmission is finished.

A. Task Model

The computational tasks generate randomly from users, which is a random process in terms of the MCC system. According to the analyzing of Google data centers, the empirical distribution of arriving intervals match negative exponential distribution [15]. Thus, we assume that the arrival of the tasks follows a Poisson process, and the total arrival rate is λ. The processing delay of each task is the total time of the offloading, which is shown in (1). If the task is executed in the edge cloud, the processing delay includes: wireless transmission delay t_{wl}, backhaul transmission delay t_b, waiting delay t_{wt} and execution delay t_{ee}. If the task is executed in the remote cloud, the processing delay includes: wireless transmission delay t_{wl}, Internet transmission delay t_I and execution delay t_{re}. In the equation, c denotes the cloud service provider that the user chooses, which is e for edge cloud and r for remote cloud.

$$t_p(c) = \begin{cases} t_{wl} + t_b + t_{wt} + t_{ee} & \text{Edge cloud execution} \\ t_{wl} + t_I + t_{re} & \text{Remote cloud execution} \end{cases} \tag{1}$$

B. Clouds Model

In the edge cloud, a buffer could hold the waiting tasks. The statistical data of Google [15] shows that the leaving intervals follow negative exponential distribution. So, we use M/M/1 queue to model the edge cloud, which is also applied in many existing papers [16][17]. For the edge cloud, assume that its service rate is μ_e, which denotes the computational resource capacity of the edge cloud. The edge cloud charges p_e for each offloading task. The cost to operate the edge cloud is $C_e(\mu_e)$, which is a function of the resource capacity μ_e.

The remote cloud has large-scale physical machines, which benefits from great pooling gain by resource aggregation. As the remote cloud has abundant computational resource, it could execute offloaded tasks right after their arrival. Thus, the remote cloud is modelled as M/M/∞ queue whose service rate is μ_r. Assume that the remote cloud charges p_r for each offloading task. Because the remote cloud has already been built, its cost is a fixed value that can not be further optimized. Thus, assume that the cost to operate the remote cloud is a given value C_r.

C. Utility Function of Users

The utility of each user is composed of two parts: the gain and the cost, which is shown in (2). Assume that the benefit of task offloading is R, which indicates the benefit obtained by the users. The QoS of the system is the user-perceived delay, which is also related to the gain of users. In this paper, we consider mean delay as the QoS of the system. The utility decreases with the increase of mean delay. The cost is the the price charged for users. Assume that η is a weighting parameter, which indicates if users emphasize more on price or delay. If the utility is smaller than a threshold U_0, users will feel like to execute the task in mobile devices rather than in the cloud [13]. As long as the threshold is reached, a user will select a cloud between the edge cloud and the remote cloud so that the utility is maximized.

$$U(c) = R - E(t_p(c)) - \eta p_c \tag{2}$$

D. Profits of CSPs

The profit of the edge cloud is composed of revenue and cost. Assume that f_e is the proportion of users who select the edge cloud, and f_r is the proportion of users who select the remote cloud. As the offloaded tasks are executed in one of the clouds, the equation $f_e + f_r = 1$ holds. Given the total arrival rate λ, the arrival rates of the edge cloud and the remote cloud are $f_e\lambda$ and $f_r\lambda$ respectively. The revenue of each cloud is related to the arrival rate and price. The profit of the edge cloud π_e and the remote cloud π_r are:

$$\begin{cases} \pi_e = f_e\lambda p_e - C_e(\mu_e) & \text{Edge cloud} \\ \pi_r = f_r\lambda p_r - C_r & \text{Remote cloud} \end{cases} \tag{3}$$

For the edge cloud, it will only be built if the revenue is a positive value, otherwise, the CSPs are not willing to provide edge cloud services. The CSPs mainly consider two parameters when building the edge cloud, which includes the computational resource capacity μ_e and the price p_e. By adjusting the two parameters, the edge cloud try to maximize its own profit while the utility threshold U_0 of users is met. The optimization problem of the edge cloud is shown as follow.

$$\max_{\mu_e, p_e} \quad \pi_{\mathrm{e}}(\mu_e, p_e) \tag{4}$$

$$s.t. \quad \pi_e > 0 \tag{5}$$

$$U(e) \geq U_0 \tag{6}$$

For the remote cloud, it also tries to maximize its profit. Considering the fact that the remote cloud has already been built, the only parameter that the remote cloud can adjust is its price p_r. As the cost to operate the remote cloud is fixed, the price p_r should not be smaller than a threshold P_{rm} so that the profit is positive. Meanwhile, as the utility of users has a lower bound, the price p_r should also not be large than a threshold $p_{rM} = \eta^{-1}(R - E(t_p(r)) - U_0)$. The optimization problem of the remote cloud is shown as follow.

$$\max_{p_r} \quad \pi_{\mathrm{r}}(p_r) \tag{7}$$

$$s.t. \quad P_{rm} \leq p_r \leq P_{rM} \tag{8}$$

III. Pricing Policy of Heterogeneous Clouds

In this section, we study the Nash equilibrium prices of CSPs when the edge computational resource capacity μ_e is given. In the Nash equilibrium prices of the market, none of the CSPs could improve the profit by solely changing its own price. Assume that (p_e^*, p_r^*) is the Nash equilibrium where p_e^* is the price of the edge cloud and p_r^* is the price of the remote cloud. In Nash equilibrium, $\forall c \in \{e, r\}$, assume that p_{-c}^* denotes the Nash equilibrium prices of CSPs other than c, $\forall p_c$ the inequality (9) holds.

$$\pi_c(p_c^*, p_{-c}^*) \geq \pi_c(p_c, p_{-c}^*) \tag{9}$$

In Nash equilibrium, each user will also not benefit from changing the choice of clouds. Thus, the utility to use the edge cloud is the same as the utility to use the remote cloud [14]. The following equation holds:

$$U(e) = U(r) \tag{10}$$

In a M/M/1 queue whose arrival rate and service rate are λ and μ respectively, the mean waiting time adds the mean execution time is $\frac{1}{\mu-\lambda}$ [18]. As the service rate of the remote cloud is μ_r, the mean execution time is $\frac{1}{\mu_r}$. Assume that the time factor $T_e = E(t_I) + \frac{1}{\mu_r} - E(t_b)$, the following equation is derived according to the equation $U(e) = U(r)$.

$$\frac{1}{\mu_e - f_e\lambda} = \eta p_r - \eta p_e + T_e \tag{11}$$

In (11), the inequality $0 \leq f_e \leq 1$ should hold, because only part of users choose the edge cloud. Thus, the following inequality holds, which is a basic condition of this work.

$$\begin{cases} \frac{1}{\mu_e} \leq \eta p_r - \eta p_e + T_e \leq \frac{1}{\mu_e - \lambda} & \text{if } \lambda < \mu_e \\ \frac{1}{\mu_e} \leq \eta p_r - \eta p_e + T_e & \text{if } \lambda \geq \mu_e \end{cases} \tag{12}$$

According to (11), the arrival rates of both edge cloud and remote cloud are derived, which are:

$$f_e\lambda = \mu_e - \frac{1}{\eta p_r - \eta p_e + T_e} \tag{13}$$

$$f_r\lambda = \lambda - \mu_e + \frac{1}{\eta p_r - \eta p_e + T_e} \tag{14}$$

Accordingly, the optimization problem of the edge cloud is:

$$\begin{aligned} \max_{\mu_e, p_e} \quad & \pi_e(\mu_e, p_e) = f_e\lambda p_e - C_e(\mu_e) \\ & = (\mu_e - \frac{1}{\eta p_r - \eta p_e + T_e})p_e - C_e(\mu_e) \end{aligned} \tag{15}$$

The optimization problem of the remote cloud is:

$$\begin{aligned} \max_{p_r} \quad & \pi_r(p_r) = f_r\lambda p_r - C_r \\ & = (\lambda - \mu_e + \frac{1}{\eta p_r - \eta p_e + T_e})p_r - C_r \end{aligned} \tag{16}$$

To get the Nash equilibrium prices, we firstly derive the partial derivation of π_e and π_r with respect to p_e and p_r:

$$\frac{\partial \pi_e}{\partial p_e} = \mu_e - \frac{p_r + \frac{T_e}{\eta}}{\eta(p_e - p_r - \frac{T_e}{\eta})^2} \tag{17}$$

$$\frac{\partial \pi_r}{\partial p_r} = \lambda - \mu_e - \frac{p_e - \frac{T_e}{\eta}}{\eta(p_e - p_r - \frac{T_e}{\eta})^2} \tag{18}$$

According to (17) and (18), the Nash equilibrium prices can be derived as long as none of the CSPs could achieve higher profit by adjusting its own price. However, the prices should also meet the constraint (12), otherwise, the stable prices are meaningless. Based on (17), the optimal price of the edge cloud is derived, which satisfy the previous constraints.

Theorem 1. Given the total arrival rate λ and the time factor T_e, the single Nash equilibrium price of the edge cloud $\hat{p}_e$ is a function of the price of remote cloud p_r and computational

capacity μ_e which is shown as follow. The price meets the constraint in (12).

$$\hat{p}_e(p_r, \mu_e) = p_r + \frac{T_e}{\eta} - \sqrt{\frac{\eta p_r + T_e}{\eta^2 \mu_e}} \tag{19}$$

Proof. Given the positive revenue constraint and (12), note that p_e and p_r satisfy the following inequality.

$$0 < p_e \leq p_r + \frac{T_e}{\eta} - \frac{1}{\eta \mu_e} \tag{20}$$

Furthermore, the following inequalities are derived.

$$\left.\frac{\partial \pi_e}{\partial p_e}\right|_{p_e=0} = \mu_e - \frac{1}{\eta p_r + T_e} \geq 0 \tag{21}$$

$$\left.\frac{\partial \pi_e}{\partial p_e}\right|_{p_e = p_r + \frac{T_e}{\eta} - \frac{1}{\eta \mu_e}} = \mu_e - \frac{p_r + \frac{T_e}{\eta}}{\frac{1}{\eta \mu_e^2}} \leq 0 \tag{22}$$

Thus, the optimal p_e exists with the constraint (12). As (17) is a monotone function, $\hat{p}_e(p_r, \mu_e)$ is derived accordingly. □

Let $\frac{\partial \pi_e}{\partial p_e} = 0$ and $\frac{\partial \pi_r}{\partial p_r} = 0$, the Nash equilibrium prices are derived with the constraint (8) and (12), and the price of remote cloud is:

$$\hat{p}_r(\mu_e) = \frac{-(2\lambda^2 - 6\lambda\mu_e + 4\mu_e^2)T_e + \mu_e + \sqrt{8\mu_e^3 T_e + \mu_e^2 - 4\lambda\mu_e T_e}}{2\eta(\lambda - 2\mu_e)^2}$$

Bringing $\hat{p}_r(\mu_e)$ into equation $\hat{p}_e(p_r, \mu_e)$, the price of the edge cloud is also got. However, the condition that $\frac{\partial \pi_r}{\partial p_r} = 0$ do not always hold with the change of μ_e, because the value of p_r is bounded. In fact, the Nash equilibrium prices are piecewise functions, which have the following several conditions.

1) If $\mu_e < (\eta P_{rM} + T_e)^{-1}$:

The edge cloud is lack of computational resource so that the users will not use it regardless of its price. In fact, $p_e \leq P_{rM} + \eta^{-1} T_e - (\eta \mu_e)^{-1} < 0$. In this case, $f_e = 0$ holds, and $p_r^* = P_{rM}, p_e^* = 0$.

2) If $\mu_e \leq \widetilde{\mu}_e$ and $\mu_e \geq (\eta P_{rM} + T_e)^{-1}$: where

$$\widetilde{\mu}_e = \frac{1}{2(\eta P_{rm} + T_e)} + \lambda + \sqrt{\frac{(\eta P_{rm} + T_e) + 4\lambda}{(\eta P_{rm} + T_e)^2}} \tag{23}$$

As $P_{rm} \leq p_r \leq P_{rM}$, if $\hat{p}_r(\mu_e) > P_{rM}, p_r^* = P_{rM}$, if $\hat{p}_r(\mu_e) < P_{rm}, p_r^* = P_{rm}$, otherwise, $p_r^* = \hat{p}_r(\mu_e)$. The Nash equilibrium price of the edge cloud is $p_e^* = \hat{p}_e(p_r^*, \mu_e)$.

3) If $\mu_e > \widetilde{\mu}_e$:

The remote cloud is not sufficient to serve mobile users because of the long delay and small weighting parameter η. In this case, $f_r = 0$ holds, and $p_r^* = P_{rm}, p_e^* = P_{rm} + \eta^{-1} T_e - \eta^{-1}(\mu_e - \lambda)^{-1}$.

IV. The Provisioning of the Computational Resource in the Edge Cloud

From the study of pricing policy, it is known that the Nash equilibrium prices of both the edge cloud and the remote cloud are functions of the resource capacity μ_e. Thus, the optimization problem of the edge cloud is as follow, and the optimization variable is μ_e.

$$\max_{\mu_e} \ \pi_e(\mu_e, p_e(p_r(\mu_e), \mu_e)) \tag{24}$$

Based on the optimization problem, the utility of the edge cloud is maximized by deploying the optimal computational capacity μ_e^*. To specify $C_e(\mu_e)$, we assume that the cost of the edge cloud is a linear function $C_e(\mu_e) = c\mu_e$.

Theorem 2. The lower bound and upper bound of the optimal computational resource capacity μ_e^* is shown as follow.

$$\mu_e^* \geq \frac{1}{\eta P_{rM} + T_e} \tag{25}$$

$$\mu_e^* \leq \max\{\frac{\eta c}{\lambda} + \lambda, \widetilde{\mu}_e\} \tag{26}$$

Proof. $\forall \mu_e$, if $\mu_e \geq (\eta P_{rM} + T_e)^{-1}$, $p_e^* = 0$ is derived according to the study of Nash equilibrium prices. As $\frac{\partial \pi_e}{\partial \mu_e} < 0$, $\pi_e < 0$ holds. Thus, $(\eta P_{rM} + T_e)^{-1}$ is the lower bound of μ_e^*.

$$\frac{\partial \pi_e}{\partial \mu_e} = \frac{\partial}{\partial \mu_e}(C_e(\mu_e)) < 0 \tag{27}$$

$\forall \mu_e$, if $\mu_e \geq \max\{\frac{\eta c}{\lambda} + \lambda, \widetilde{\mu}_e\}$, $p_r^* = P_{rm}$ and $p_e^* = P_{rm} + \eta^{-1} T_e - \eta^{-1}(\mu_e - \lambda)^{-1}$ are derived according to the study of Nash equilibrium prices. As $\frac{\partial \pi_e}{\partial \mu_e} < 0$, the upper bound of μ_e^* is derived.

$$\frac{\partial \pi_e}{\partial \mu_e} = \frac{\partial}{\partial \mu_e}(-\frac{\lambda}{\eta(\mu_e - \lambda)} - c\mu_e) < 0 \tag{28}$$

□

Based on the Theorem 2, it is noted that the optimal computational resource capacity has a lower bound and a upper bound. Thus, the optimization problem is to find the μ_e^* that maximize π_e within a given range. However, the $\pi_e(\mu_e)$ has more than one local optimal values, and these values are hard to be expressed by formulations. To solve the optimization problem, we design a search algorithm which is shown in Algorithm 1.

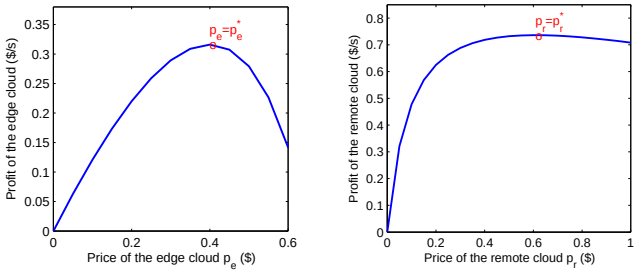
V. Numerical Results

In this section, numerical results are shown to validate the previous study on the Nash equilibrium prices and the provisioning of computational resource. To specify the services provided by CSPs, we consider face recognition applications which can be either executed in the remote cloud or in the edge cloud. For the parameters of delay, we rely on some existing works to get the representative values. The Internet transmission delay has very large variance, and the mean value is assumed to be 100 to 300 milliseconds [19]. The execution delay is about 500 milliseconds [20]. The data transmission delay between the edge cloud and the base station is assumed to be 10 milliseconds.

Algorithm 1 Find the optimal computational resource capacity of the edge cloud

Input: $\lambda, \eta, \mu_r, E(t_I), E(t_b), P_{rm}, P_{rM}, c, \Delta\mu$
Output: μ_e^*

$T_e \leftarrow E(t_I) + \frac{1}{\mu_r} - E(t_b)$
$\mu_{min} \leftarrow \frac{1}{\eta P_{rM} + T_e}$
$\mu_{max} \leftarrow \max\{\frac{\eta c}{\lambda} + \lambda, \frac{1}{2(\eta P_{rm} + T_e)} + \lambda + \sqrt{\frac{\eta P_{rm} + T_e + 4\lambda}{(\eta P_{rm} + T_e)^2}}\}$
$\mu_e^* \leftarrow \mu_{min}$
$\pi_e^* \leftarrow 0$
$\mu_e \leftarrow \mu_{min}$
while $\mu_e \leq \mu_{max}$ **do**
  if $\hat{p}_r(\mu_e) > P_{rM}$ **then**
    $p_r^* \leftarrow P_{rM}$
  else if $\hat{p}_r(\mu_e) < P_{rm}$ **then**
    $p_r^* \leftarrow P_{rm}$
  else
    $p_r^* \leftarrow \hat{p}_r(\mu_e)$
  end if
  $p_e^* \leftarrow \hat{p}_e(p_r^*, \mu_e)$
  if $\pi_e(\mu_e) > \pi_e^*$ **then**
    $\pi_e^* \leftarrow \pi_e(\mu_e)$
    $\mu_e^* \leftarrow \mu_e$
  end if
  $\mu_e \leftarrow \mu_e + \Delta\mu$
end while
if $\pi_e^* < 0$ **then**
  $\mu_e^* \leftarrow 0$
end if

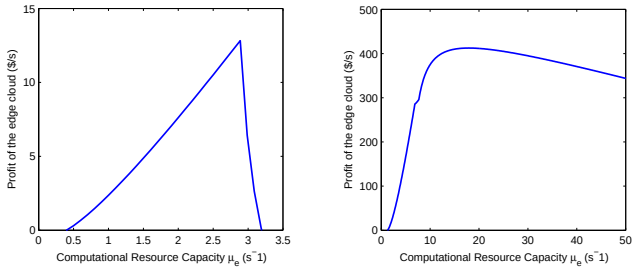


(a) Edge cloud ($p_r = p_r^*$) (b) Remote Cloud ($p_e = p_e^*$)

Fig. 2. The Nash equilibrium prices

Fig. 2. shows the Nash equilibrium prices of both the edge cloud and the remote cloud. The prices of the remote cloud and the edge cloud are p_r^* and p_e^* respectively. The left figure is the edge cloud profit with fixed remote cloud price $p_r = p_r^*$, and the x-axis is the price of edge cloud p_e. The p_e that maximize the profit is p_e^*. Similarly, The right figure is the remote cloud profit with fixed edge cloud price $p_e = p_e^*$, and p_r that maximize the profit is p_r^*. These two figures indicate that none of the CSPs could improve its profit by changing the price solely in the Nash equilibrium prices.

Fig. 3. shows the relationship between the computational resource capacity of the edge cloud and its profit. In fact, the



(a) $\eta = 1$ (b) $\eta = 0.01$

Fig. 3. Edge cloud utility vs. Computational resource capacity

profit is highly related to the weighting parameter η. The left figure is the case that $\eta = 1$ in which a turning point appears in the function. In fact, η is a relatively large value so that users emphasize more on price than on delay. When remote cloud begins to decrease the price, users are more willing to choose it even though the delay is large. Thus, the profit of the edge cloud decreases rapidly when μ_e reaches a threshold where the remote cloud begins to adjust its price, and this is the turning point. The right figure is the case that $\eta = 0.01$ where users emphasize on the delay much more than the price. Even though the remote cloud adjusts its price in some conditions, it weakly influences the profit of the edge cloud. This is because that uses mainly consider the processing delay, and the decrease of the remote cloud price do not change the selection of them.

Fig. 4. shows the relationship between the optimal computational resource capacity of the edge cloud μ_e^*, the arrival rate of users λ and the weighting parameter η. With the increase of the arrival rate, the μ_e^* is nearly growing linearly. This trend indicates the linear relationship between the demand of users and the supply of resource. Furthermore, the η is an important parameter which determines the μ_e^*. When η becomes smaller, users emphasize more on delay than on price. In this case, more computational resource should be deployed locally so that users could have their delay requirements met.

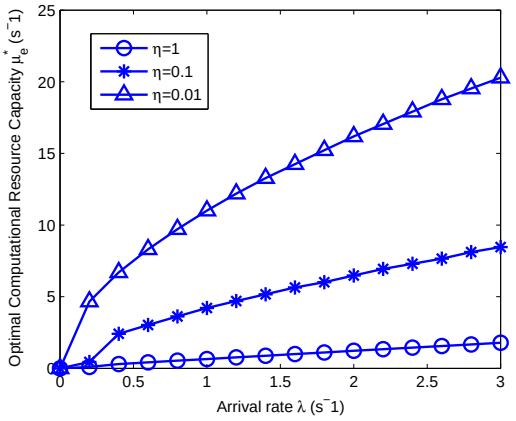


Fig. 4. Optimal computational resource capacity vs. Arrival rate.

VI. Conclusions

In this paper, we study the provisioning of edge computational resource to maximize the profit of the edge cloud. As the edge cloud and the remote cloud provide similar computational services to mobile users, their prices deeply influence the profit of each other. Meanwhile, the profit of the edge cloud is also highly related to the provisioning cost. To solve this optimization problem, we firstly derive the Nash equilibrium prices of heterogeneous clouds according to the computational resource capacity of the edge cloud. Based on the Nash equilibrium prices, we get the theoretical lower and upper bound of the optimal computational resource capacity. We finally design an algorithm to get the optimal computational resource capacity of the edge cloud. Numerical results show that the optimal computational resource capacity of the edge cloud is nearly a linear function of the arrival rate of users. Meanwhile, the resource capacity is highly related to a weighting parameter η which indicates if users emphasize more on price or delay. The price of the remote cloud greatly influences the profit of the edge cloud when users care more about price. However, when users care more about delay, the price of the remote cloud weakly influences the profit of the edge cloud.

Acknowledgment

This work is sponsored in part by the National Basic Research Program of China (973 Program: No. 2012CB316001), the National Science Foundation of China (NSFC) under grant No. 61201191, No. 61322111, No. 61321061, No. 61401250, and No. 61461136004, and Intel Collaborative Research Institute for Mobile Networking and Computing.

References

[1] K. Kumar and Y. H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *Computer*, 2010.

[2] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Transactions on Wireless Communications*, 2013.

[3] J. Aikat, J. Kaur, F. D. Smith, and K. Jeffay, "Variability in tcp round-trip times," in *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, 2003.

[4] M. Claypool and K. Claypool, "Latency and player actions in online games," *Communications of the ACM*, 2006.

[5] M. T. Beck, S. Feld, C. Linnhoff-Popien, and U. Pützschler, "Mobile edge computing," *Informatik-Spektrum*, 2016.

[6] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *IEEE Pervasive Computing*, 2009.

[7] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5g heterogeneous networks," *IEEE Signal Processing Magazine*, 2014.

[8] J. Liu, T. Zhao, S. Zhou, Y. Cheng, and Z. Niu, "Concert: a cloud-based architecture for next-generation cellular systems," *IEEE Wireless Communications*, 2014.

[9] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys Tutorials*, 2016.

[10] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM computer communication review*, 2008.

[11] R. Harms and M. Yamartino, "The economics of the cloud," *Microsoft whitepaper, Microsoft Corporation*, 2010.

[12] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "On the statistical multiplexing gain of virtual base station pools," in *Global Communications Conference (GLOBECOM), 2014 IEEE*, 2014.

[13] Y. Feng, B. Li, and B. Li, "Price competition in an oligopoly market with multiple iaas cloud providers," *IEEE Transactions on Computers*, 2014.

[14] X. Li, B. Gu, C. Zhang, K. Yamori, and Y. Tanaka, "Price competition in a duopoly iaas cloud market," in *Network Operations and Management Symposium (APNOMS), 2014 16th Asia-Pacific*, 2014.

[15] C. Jiang, Y. Chen, Q. Wang, and K. J. R. Liu, "Data-driven stochastic scheduling and dynamic auction in iaas," in *2015 IEEE Global Communications Conference (GLOBECOM)*, 2014.

[16] M. Guevara, B. Lubin, and B. C. Lee, "Navigating heterogeneous processors with market mechanisms," in *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on*, 2013.

[17] J. Bi, Z. Zhu, R. Tian, and Q. Wang, "Dynamic provisioning modeling for virtualized multi-tier applications in cloud data center," in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, 2010.

[18] D. Gross, *Fundamentals of Queueing Theory*, 2008.

[19] M. DeVirgilio, W. D. Pan, L. L. Joiner, and D. Wu, "Internet delay statistics: Measuring internet feel using a dichotomous hurst parameter," in *Southeastcon, 2013 Proceedings of IEEE*, 2013.

[20] N. Powers, A. Alling, K. Osolinsky, T. Soyata, M. Zhu, H. Wang, H. Ba, W. Heinzelman, J. Shi, and M. Kwon, "The cloudlet accelerator: Bringing mobile-cloud face recognition into real-time," in *2015 IEEE Globecom Workshops (GC Wkshps)*, 2015.