Price Competition in an Oligopoly Market with Multiple IaaS Cloud Providers

Yuan Feng*, Baochun Li†, Bo Li‡
*Department of Computing, Hong Kong Polytechnic University
†Department of Electrical and Computer Engineering, University of Toronto
‡Department of Computer Science, Hong Kong University of Science and Technology

***Abstract*—As an increasing number of Infrastructure-as-a-Service (IaaS) cloud providers start to provide cloud computing services, they form a *competition market* to compete for users of these services. Due to different resource capacities and service workloads, users may observe different finishing times for their cloud computing tasks and experience different levels of service qualities as a result. In order to compete for cloud users, it is critically important for each cloud service provider to select an "optimal" price that best corresponds to their service qualities, yet remaining attractive to cloud users. To achieve this goal, the the underlying rationale and characteristics in this competition market need to be better understood.**

In this paper, we present an in-depth game theoretic study of such a competition market with multiple competing IaaS cloud providers. We characterize the nature of non-cooperative competition in an IaaS cloud market, with a goal of capturing how each IaaS cloud provider will select its optimal prices to compete with the others. Our analyses lead to sufficient conditions for the existence of a Nash equilibrium, and we characterize the equilibrium analytically in special cases. Based on our analyses, we propose iterative algorithms for IaaS cloud providers to compute equilibrium prices, which converge quickly in our study.

***Index Terms*—Cloud computing, Infrastructure-as-a-Service, market competition, cloud pricing**

I. Introduction

Cloud computing has recently emerged as a new paradigm for a cloud provider to host and deliver computing services to enterprises and consumers who use such services. One of the possible types of cloud services provided by today's cloud providers, such as Amazon EC2 and Rackspace, is referred to as Infrastructure as a Service (IaaS). With IaaS, each physical machine that a cloud provider hosts is virtualized using a hypervisor, such as Xen Server. Such virtualization makes it feasible for each physical machine to host multiple virtual machines (VMs), and computing resources are leased to cloud users in the form of these VMs. By migrating from traditional in-house server infrastructures to cloud computing, cloud users may trade a significant amount of up-front investment costs to the ongoing costs of using resources provisioned on-demand by IaaS cloud providers. In return, IaaS cloud providers are able to charge their users for using computing resources on a "pay-as-you-go" basis.

With multiple IaaS cloud providers available to the cloud users, one of the ways they may differentiate themselves with is their pay-as-you-go prices of using VMs for an hour, and such prices reflect the quality of their services. For example, as of March 2013, for each Virtual Machine (VM) with 4 CPU cores, Amazon EC2 charges $\$0.24$ for an hour of usage (called *large on-demand instance*) [1], and GoGrid charges $\$0.32$ [2], and Rackspace charges $\$0.48$ [3].

Since a user's cloud service demand may be satisfied by any of these IaaS cloud providers, a rational user will choose the one that maximizes its own net reward, *i.e.*, its utility obtained by choosing the IaaS cloud service minus its payment. The utility of a user is not only determined by the *importance* of the task (*i.e.*, how much benefit the user can receive by finishing this task), but also closely related to the *urgency* of the task (*i.e.*, how quickly it can be finished). The same task, such as running an online voice recognition algorithm, is able to generate more utility for a cloud user if it can be completed within a shorter period of time in the cloud. Since the diversity among IaaS cloud providers will lead to different net rewards, multiple IaaS providers form a market to *compete* for cloud users.

Existing real-world measurement results [4] reveal that different IaaS providers complete tasks with different completion times, and an IaaS provider can become less competitive with an inappropriate price setting. With different price settings, payments made to finish each benchmarking task are also different across different providers. As a consequence, the IaaS cloud providers are presented with a question: how can each provider compute the *optimal* price to maximize its profit in such a competition market, in which demands from cloud users are sensitive to both the finishing time and the payment of completing a task?

It turns out that answering this question is non-trivial. On one hand, IaaS cloud providers may wish to increase the price to generate more profit. On the other hand, increasing the price too much in a competitive environment may risk losing potential cloud users, which then results in a reduced amount of profit. Further, although reducing the price should intuitively be an effective way to attract cloud users, these users may overwhelm the IaaS cloud provider due to an unreasonably low price, which then leads to longer finishing times on the tasks to be completed. As a consequence, the reduced utility will prohibit future users to choose this cloud provider.

In this paper, we take the first step to study the price competition in a cloud market formed by multiple IaaS cloud providers. More specifically, we present an in-depth analytical study on the monopoly, duopoly and oligopoly markets, in which multiple IaaS cloud providers are competing with one another. We use an M/M/1 queue to model correlations among the expected task finishing times, an IaaS cloud provider's resource capacity, and the request rates from cloud users. Since the pricing strategy of a cloud provider depends on its competitors, we take a game-theoretic perspective to study the strategic situation. To our knowledge, this is the first study that discusses the competition among IaaS cloud providers in the context of oligopoly market competition.

Our original contributions in this paper hinge upon the sufficient conditions we have derived for the existence of a Nash equilibrium in the market. By analyzing the Nash equilibrium, we make the following observations. *First*, when multiple IaaS cloud providers compete for users, the cloud provider with a larger resource capacity is able to charge a higher price and take more cloud users in equilibrium. However, its profit will not monotonically increase with larger resource capacities, due to increasing operating costs. As a result, though increasing the resource capacity is an effective way for a cloud provider to become more competitive in the market, it can only increase its expected profit to a certain extent. If we take service-level objectives, security measures, reputation and brand into consideration, increasing the capacity of a datacenter may become even less effective. *Second*, the equilibrium price is found to be sensitive to the importance as well as the urgency of tasks of cloud users: it decreases with the importance and increases with the urgency of tasks. This motivates the use of service-level objectives for cloud users to further specify the importance and urgency of their tasks. *Third*, the equilibrium prices are not always socially optimal. *Finally*, we propose iterative algorithms to find equilibrium prices in the duopoly and oligopoly markets, respectively, both of which are shown to be converging rapidly to the equilibrium.

The remainder of this paper is organized as follows. We show the originality of our work in the context of related work in Sec. II. In Sec. III, we first formulate the competition market and present our model, and then begin our analysis with the monopoly problem, which serves as the baseline for our comparisons. In Sec. IV, we analyze the competition between two IaaS cloud providers with heterogeneous users, and propose an iterative algorithm to find equilibrium prices. We also study the corresponding social welfare problem. We extend our discussion to an oligopoly market in Sec. IV-B, and propose an algorithm to find Nash equilibrium prices for each cloud provider. Sec. V shows some characteristics of Nash equilibrium prices with extensive simulations. In Sec. VI, we conclude the paper with extensive discussions on other important factors that influence the pricing strategies in a cloud market.

II. Related Work

Considerable performance differences across cloud providers have attracted a substantial amount of research attention. Hong *et al.* [5] and Tsakalozos *et al.* [6] applied dynamic programming and microeconomics, respectively, to achieve optimal resource allocation for cloud users in VM-based IaaS clouds, with full awareness of different prices charged by cloud providers.

Existing papers were concerned with the problem of how optimal pricing in the cloud can be achieved. To find the optimal price for a caching service in the cloud, Kantere *et al.* [7] modeled the correlation between user demand and the price, and proposed a dynamic pricing scheme to maximize the cloud provider's profit. Teng *et al.* [8] and Mihailescu *et al.* [9] studied optimal pricing with an auction mechanism, in which users had budgetary and deadline constraints. Our previous work [10] considered an exchange-based market for VMs, and proposed a solution based on Nash bargaining games. Xu *et al.* [11] used a revenue management framework to maximize a cloud provider's revenue with dynamic cloud pricing. Our work in this paper differs substantially from previous papers. *First*, all previous works considered the pricing of one provider alone, but our focus in this paper is how optimal pricing can be determined in a competitive environment with more than one cloud provider. *Second*, most previous models assume that the price is a certain function of user demand, which has not been validated in measurement studies. In contrast, we make the more realistic assumption that user demand at each cloud provider remains unknown, and is subject to a game-theoretic analysis in a duopoly or oligopoly cloud market.

Price competition has been an active research topic in the context of economic markets with multiple service providers. Petri *et al.* [12], [13] have studied the effects of risk in service-level agreements (SLAs) in service provider communities. Chen *et al.* have presented an analysis of the equilibrium price in a monopoly market [14], and they have also discussed equilibrium prices in a duopoly market with varying demand [15]. Allon *et al.* examined the scenario that multiple providers competed for users using different prices and time guarantees [16]. The competition game among multiple resource providers was also considered in networking research. Anselmi *et al.* studied a congestion game with multiple links, each of which was under the control of a profit maximizing provider [17]. In the context of processor sharing queues, they discussed the existence and efficiency of oligopolistic equilibria.

Similar to these existing works, we are also interested in the existence of Nash equilibria in the cloud market with multiple IaaS cloud providers. Yet, the context of our study is price competition in a cloud computing environment, which has a different system model. In our model, each cloud user is associated with a different request rate as it is served by the cloud, and such heterogeneity in per-user request rates makes our analyses much more challenging.

III. Model Formulation and Monopoly Analysis

To begin with, we present our system model in the context of IaaS cloud providers, and establish important results with respect to monopoly pricing, which, while being the most elementary in our analyses, provides us with a solid understanding towards our main analytical results that follow.

A. System Model

In this paper, we are concerned with a market with multiple IaaS cloud providers, who are competing for cloud users. Each cloud provider is modeled by an M/M/1 queue, serving a common pool of potential cloud users with one "super" server, which combines the resource capacity of multiple physical

TABLE I
DEFINITIONS OF MATHEMATICAL NOTATIONS

Notation	Definition
μ_i	the service rate of the cloud provider i
$\gamma_i(\mu_i)$	the operating cost at the cloud provider i
p_i	the usage price per VM at the cloud provider i
f_i	the market share of the cloud provider i
v	the reservation value at cloud users
M	the number of cloud users
λ_j	the request rate at cloud user j
$U_i(\lambda_j)$	the utility of cloud user j by choosing to be served by the IaaS provider i
r	the benefit factor per VM requested
c	the cost factor per time unit
π_i	the expected profit of cloud provider i
Λ	the market size
P_{ij}	the total payment user j makes to cloud provider i
L_i	the combined effects of other competitive factors
t_i	the expected finishing time of a unit request experienced at the cloud provider i
β_i	the combined attraction of provider i's competitors

servers that the provider manages. When it comes to analyzing the response time exhibited when processing requests as a function of the computational capacity and the request arrival rate, the M/M/1 queuing model has been adopted by a number of existing papers in the literature that analyzed datacenter operations [18]–[21]. The *resource capacity* of each cloud provider i is represented by its service rate μ_i. $\gamma_i(\mu_i)$ is used to denote the operating cost at cloud provider i, which is assumed to be a function of its resource capacity. For users who would like to choose the cloud service, the IaaS cloud provider will charge a fixed per-time-unit usage price for each type of resources consumed to finish their tasks.

All operational IaaS cloud providers support *on-demand* pricing for users to use cloud computing resources. On-demand pricing allows users to pay for the amount of resources consumed to complete their tasks with no long-term commitments. With this pricing scheme, cloud providers charge users based on the amount of resources consumed to complete their tasks. As a result, we use p_i^r to denote the fixed usage price per resource unit — for example, an unit of CPU time when using a virtual machine — at an IaaS provider i for a type of resource r. As we will focus on the price competition among multiple IaaS providers for a given type of resource, the indices r will be dropped for simplicity.

The arrival of requests from cloud users are assumed to follow a Poisson process, an assumption that is commonly used in competition models in the economic literature [14]–[16]. A cloud user j makes a choice to be served by a specific cloud provider. Yet, it also maintains a *reservation value* v (assumed to be the same across all users), and if by using the cloud service its net reward falls below v, user j can refuse to use any cloud service, and choose to finish its task locally.

As shown in Fig. 1, a user j has a task with requests for resources that it wishes to finish in the cloud. The rate at which these requests are generated when running the task at a cloud provider is denoted by λ_j. The *market share* of a cloud provider i is denoted by f_i, which equals the sum of request rates of all users who choose cloud provider i. Each cloud user

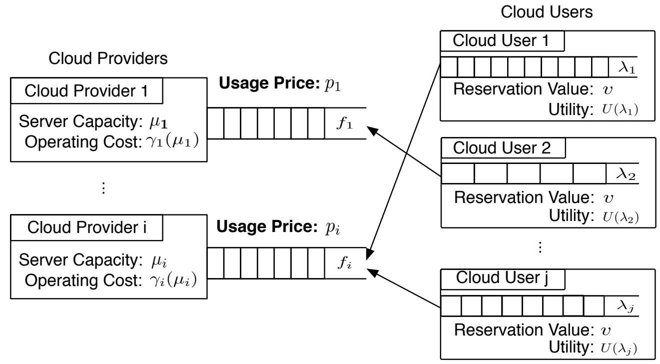


Fig. 1. Our model of competition in an oligopoly cloud market with multiple IaaS cloud providers.

only selects one of the IaaS cloud providers, *i.e.*, it does not split its requests by routing them to multiple IaaS providers simultaneously.

Since the cloud provider i is modeled as an M/M/1 queue with a service rate μ_i, based on queueing theory [22], the expected finishing time experienced by a request from one of the cloud users (called *response time* in the queueing theory literature), including both the time waiting in the queue and the time being served, is $\frac{1}{\mu_i - f_i}$. This does not depend on the scheduling discipline as long as it is work conserving, and can be computed using Little's Law. It then follows that the expected time that a user request spent waiting in the queue is $\frac{1}{\mu_i - f_i} - \frac{1}{\mu_i}$, the second term being the expected service time. With a request rate of λ_j at cloud provider i, the expected finishing time is, therefore:

$$\frac{1}{\mu_i - f_i} - \frac{1}{\mu_i} + \frac{\lambda_j}{\mu_i} = \frac{f_i}{\mu_i(\mu_i - f_i)} + \frac{\lambda_j}{\mu_i}. \tag{1}$$

We focus on stationary analysis only, in that for a given μ_i, the equilibrium market share $f_i < \mu_i$ must hold, since otherwise the queue will grow infinitely long and the system will not have a stationary distribution. For the problem to be worth discussing, we assume $f_i < \mu_i$ throughout the entire paper.

Now let us consider user j's utility, by choosing to be served by a cloud provider i. Since a user will be more satisfied with more tasks completed or faster service in the cloud, its utility should become higher as the request rate λ_j increases, and as the expected task finishing time becomes shorter. More formally, with a request rate of λ_j, user j's utility is:

$$U_i(\lambda_j) = r \cdot \lambda_j - c \cdot \left[\frac{f_i}{\mu_i(\mu_i - f_i)} + \frac{\lambda_j}{\mu_i} \right], \tag{2}$$

where the benefit factor r (per unit of resource requested) reflects the relative importance of the task, and the waiting cost factor c (per time unit) reflects its urgency. The more important a task is, the more utility is brought to the user. If a task is more urgent, *i.e.*, associated with a larger c, the time consumed to complete the task incurs a higher waiting cost.

In Fig. 1, it can be observed that multiple cloud users with distinct request rates and utilities are making a choice of their respective IaaS cloud providers, based on their own utilities and on the prices charged by each cloud provider. Each IaaS cloud provider sets an appropriate usage price, according to their resource capacities and the prices from the other providers. The objective of each cloud provider is to maximize its own expected profit, with full awareness of possible reactions of other cloud providers and all the cloud users. Important notations that we introduced so far are presented in Table I.

Remark. In some existing papers on the topic of analyzing cloud computing performance, *e.g.,* [23]–[25], a cloud provider is modeled as a M/M/c queue. This model is apparently motivated in the literature by the fact that an IaaS cloud provider operates multiple physical servers, and each physical server can be viewed as a *server* from a queueing theoretic perspective in the M/M/c model. If we adopt this model, the resource capacity of each cloud provider i can then be represented by its service rate $\mu_i = n_i\mu_i'$, where the cloud provider i has n_i servers, each of which has a service rate of μ_i'.

Based on the theory of M/M/c queues [22], the expected finishing time experienced by a user with a request rate of λ_j at cloud provider i is $\frac{p_q}{n_i\mu_i'-f_i} - \frac{1}{\mu_i'} + \frac{\lambda_j}{\mu_i'}$, where p_q is the probability that an arriving request from a user will be forced to join and wait in the queue (all servers are occupied), and is given by $C(n_i, f_i/\mu_i')$, called *Erlang's C formula*. Due to the large number of physical servers in a typical IaaS cloud provider, we believe that it is sufficient to abstract multiple servers as a single "super" server to serve the market in the context of cloud providers, and the additional modeling complexity is not necessary.

B. An Analysis of Monopoly Pricing

We are now ready to present our analytical results on monopoly pricing, which serve as preliminaries and a basis for later comparisons. In this subsection, we consider a single cloud provider modeled by an M/M/1 queue, with a service rate μ and an operating cost $\gamma(\mu)$.

A rational cloud user j will seek to maximize its expected net reward by finishing the task, *i.e.*, its utility obtained by choosing the cloud service minus its total payment. Since cloud users are charged based on how much resource they consume, cloud user j's total payment $P_j = p\lambda_j, \forall j$. Now that there is only one cloud provider in the market, this implies that the cloud user will choose to use the cloud service if $U(\lambda_j) - P_j \geq v$ and refuse to use it otherwise. In equilibrium, $U(\lambda_j) - P_j = v$, with the corresponding market share $f < \mu$. In equilibrium, this implies:

$$r\lambda_j - c\left[\frac{f}{\mu(\mu-f)} + \frac{\lambda_j}{\mu}\right] - p\lambda_j = v, \forall j, \tag{3}$$

where $f = \sum_j \lambda_j$, representing the sum of request rates of all users who choose this cloud provider. Considering all cloud users together, Eqn. (3) is equivalent to

$$f = \frac{Mv\mu(\mu-f) + Mcf}{(r\mu - c - p\mu)(\mu - f)}, \tag{4}$$

if there are M cloud users.

The cloud provider's problem is to maximize its expected profit, denoted by π. That is,

$$\begin{aligned} \max_{p\geq 0} \quad & \pi = pf - \gamma(\mu) \\ \text{s.t.} \quad & f = \frac{Mv\mu(\mu-f) + Mcf}{(r\mu - c - p\mu)(\mu - f)} \\ & \mu > f \geq 0. \end{aligned} \tag{5}$$

It is worth noting that, although both the optimal price and the market share of the cloud provider is assumed to be unknown in this problem, the optimization variable is the price only. The reason is that the market share of the cloud provider is not an independent variable, and it is a function of the price in essence. Once the price of the cloud provider is chosen, cloud users will make their decisions of whether to choose this cloud service or not based on their reservation values of the tasks. This implies that as long as the price is determined, the market share is a deterministic value.

Solving the optimization problem (5), the optimal usage price for the cloud provider is found to be $p^* = \max\{p_m, p_\Lambda\}$, where Λ is referred to as the market size and equals the sum of request rates of all cloud users in the market. In the monopoly market, Λ equals to the market share f if all users choose to use the cloud provider. When the market size $\Lambda > \mu - \sqrt{\frac{\mu Mc}{\mu r - c}}$, the cloud provider is not able to take the entire market in equilibrium. The equilibrium price p^* equals the first-order price, which takes the form of:

$$p_m = r - \frac{c}{\mu} - \frac{1}{\mu}\sqrt{\frac{Mc(r\mu - c)}{\mu}} - \frac{Mv\sqrt{r\mu - c}}{\mu\sqrt{r\mu - c} - \sqrt{Mc\mu}}, \tag{6}$$

with the corresponding market share $f^* = \mu - \sqrt{\frac{\mu Mc}{\mu r - c}}$. Otherwise, the cloud provider will serve all cloud users, *i.e.*, $f^* = \Lambda$, and the optimal price p^* is:

$$p_\Lambda = r - \frac{c}{\mu} - \frac{Mc}{\mu(\mu - \Lambda)} - \frac{Mv}{\Lambda}. \tag{7}$$

In summary, when there is only one cloud provider in the market, there exists a unique optimal price $p^* = \max\{p_m, p_\Lambda\}$. The optimal market share is bounded by the cloud provider's resource capacity. If its capacity is large enough, then the cloud provider can take the entire market. Due to the existence of the reservation value, the cloud provider can not increase the price without bounds, even when it is the only provider in the market.

IV. Price Competition among Multiple IaaS Cloud Providers

A. The Duopoly Case

As a starting point, we first consider the case of a duopoly cloud market, in which two IaaS cloud providers compete with each other, with a similar game theoretic analysis as the monopoly case. In this context, we derive the relationship between the equilibrium prices for each cloud provider, and analyze the comparative statics of Nash equilibrium prices.

We first discuss how decisions are made by cloud users in this market. All cloud users act in a selfish fashion so as to maximize their own expected net reward. The optimal choice of cloud user j is to choose the cloud provider i from which it obtains a maximized net reward, or to refuse to use the cloud service if its net reward failed to exceed its reservation value. That is, a cloud user j will choose a cloud provider i (or the option of choosing neither cloud provider) that achieves

$$\max\{U_i(\lambda_j) - P_j, v\},\ i = 1, 2. \tag{8}$$

1) Nash Equilibrium in a Duopoly Market: Let π_i be the expected profit of cloud provider i. Each cloud provider i seeks to maximize π_i by choosing its usage price p_i, which clearly depends on the reaction of the other cloud provider and that of all cloud users. Let $\pi_i(p_1, p_2)$ denote the expected profit of cloud provider i if it chooses a price p_i given the other cloud provider k's price p_k, $i \neq k$ and $i, k = 1, 2$. A pair of prices (p_1^*, p_2^*) is said to be a Nash equilibrium if it satisfies:

$$\begin{aligned} \pi_1(p_1^*, p_2^*) &\geq \pi_1(p_1, p_2^*),\ \forall p_1 \geq 0, \\ \pi_2(p_1^*, p_2^*) &\geq \pi_2(p_1^*, p_2),\ \forall p_2 \geq 0. \end{aligned}$$

In a Nash equilibrium, any cloud provider can not increase the expected profit by changing its price unilaterally. That is equivalent to say, the Nash equilibrium price is the optimal price a cloud provider can achieve in a market when cloud providers do not cooperate with each other. In the equilibrium, the expected profits of both cloud providers are maximized, and the market is balanced dynamically. In our subsequent analysis, we aim to prove whether such equilibrium exists in the duopoly market, and how can each cloud provider achieve the equilibrium price if it exists.

The equilibrium prices can be found by a standard procedure of identifying the best response function of each cloud provider. Let $p_i = F_i(p_k)$ be cloud provider i's optimal price given the usage price p_k selected by cloud provider k. A Nash equilibrium in this duopoly competition market is then a pair of prices (p_1, p_2) such that $p_1 = F_1(p_2)$ and $p_2 = F_2(p_1)$, *i.e.*, an intersecting point of two best response functions.

Take cloud provider 1 as an example. The best response function F_1 can be found by assuming that cloud provider 2's price p_2 is given and by solving cloud provider 1's problem as follows:

$$\max_{p_1 \geq 0} \quad \pi_1 = \sum_j P_{1j} - \gamma_1(\mu_1) \tag{9}$$

$$\text{s.t.} \quad U_1(\lambda_j) - P_{1j} \geq v_j,\ \forall j \tag{10}$$

$$U_1(\lambda_j) - P_{1j} = U_2(\lambda_j) - P_{2j},\ \forall j \tag{11}$$

$$f_1 + f_2 = \sum_j \lambda_j \leq \Lambda$$

$$\mu_1 > f_1 \geq 0,$$

where P_{ij} is the total payment user j makes to cloud provider i. Both constraints (10) and (11) come from optimizing cloud users' net rewards. Constraint (10) indicates that for any user to choose cloud provider 1, it should be offered at least the same expected net reward as its reservation value of the task. If this constraint does not hold, the cloud user would prefer to finish its task locally rather than using the cloud service. Constraint (11) states that in equilibrium, the expected net rewards that a cloud user can derive from different cloud providers should be the same, which prohibits any cloud user from switching cloud providers.

Similarly, the optimal price of cloud provider 2 can be found by solving its corresponding problem, under the assumption that the price of cloud provider 1, p_1, is given.

$$\begin{aligned} \max_{p_2 \geq 0} \quad & \pi_2 = \sum_j P_{2j} - \gamma_2(\mu_2) \\ \text{s.t.} \quad & U_2(\lambda_j) - P_{2j} \geq v_j,\ \forall j \\ & U_1(\lambda_j) - P_{1j} = U_2(\lambda_j) - P_{2j},\ \forall j \\ & f_1 + f_2 = \sum_j \lambda_j \leq \Lambda \\ & \mu_2 > f_2 \geq 0, \end{aligned}$$

Each cloud provider will update its prices with respect to the reaction of its competitor and all cloud users, until an equilibrium point is reached, *i.e.*, when neither cloud provider can gain a higher expected profit by changing its own price unilaterally.

When cloud users are charged based on their usage of resources, the problem of finding cloud provider 1's best response function (9) is equivalent to:

$$\begin{aligned} \max_{p_1 \geq 0} \quad & \pi_1 = p_1 f_1 - \gamma_1(\mu_1) \\ \text{s.t.} \quad & f_1 \geq \frac{Mv\mu_1\delta_1 + Mcf_1}{(r\mu_1 - p_1\mu_1 - c)\delta_1} \\ & f_1 + f_2 = \frac{cf_1\mu_2\delta_2 - cf_2\mu_1\delta_1}{[c(\mu_1 - \mu_2) - \mu_1\mu_2(p_1 - p_2)]\delta_1\delta_2} \\ & f_1 + f_2 \leq \Lambda \\ & \mu_1 > f_1 \geq 0, \end{aligned}$$

where $\delta_i = \mu_i - f_i$.

By considering the best response problems of both cloud providers together, we derive the necessary condition for the existence of a Nash equilibrium. Any equilibrium must satisfy the following constraints, referred to as the first-order necessary condition for the existence of a Nash equilibrium, as summarized in Lemma 1.

***Lemma** 1:* The necessary condition for a set of solution (p_1, p_2, f_1, f_2) to be a Nash equilibrium is that, it should satisfy the following constraints:

$$f_i \geq \frac{Z_i}{Y_i\mu_i\delta_i} \tag{12}$$

$$f_1 + f_2 = \frac{cQ}{X\delta_1\delta_2} \tag{13}$$

$$f_1 + f_2 = \Lambda \tag{14}$$

$$f_1 \geq \frac{Q\delta_2(p_1 Z_1\delta_1 + Y_1^2 f_1\delta_1^2 - Mc\mu_1 f_1 Y_1)}{QY_1\delta_1\delta_2(Y_1\delta_1 - Mc\mu_1) + XZ_1(\delta_1^2 + \delta_2^2)} \tag{15}$$

$$f_2 \geq \frac{Q\delta_1(p_2 Z_2\delta_2 + Y_2^2 f_2\delta_2^2 - Mc\mu_2 f_2 Y_2)}{QY_2\delta_1\delta_2(Y_2\delta_2 - Mc\mu_2) + XZ_2(\delta_1^2 + \delta_2^2)} \tag{16}$$

$$\frac{X(X\delta_2^2 + c\mu_1\mu_2)(p_1^r Z_1\delta_1 + Y_1^2 f_1\delta_1^2 - Mc\mu_1 f_1 Y_1)}{QY_1\delta_1\delta_2(Y_1\delta_1 - Mc\mu_1) + XZ_1(\delta_1^2 + \delta_2^2)} \geq 0 \tag{17}$$

$$\frac{X(c\mu_1\mu_2 - X\delta_1^2)(p_2^r Z_2\delta_2 + Y_2^2 f_2\delta_2^2 - Mc\mu_2 f_2 Y_2)}{QY_2\delta_1\delta_2(Y_2\delta_2 - Mc\mu_2) + XZ_2(\delta_1^2 + \delta_2^2)} \geq 0, \tag{18}$$

where $X = c(\mu_1 - \mu_2) - \mu_1\mu_2(p_1 - p_2)$, $Y_i = r\mu_i - p_i\mu_i - c$, $Z_i = Mv\mu_i^2(\mu_i - f_i) + Mv\mu_i f_i$, and $Q = f_1\mu_2(\mu_2 - f_2) - f_2\mu_1(\mu_1 - f_1)$.

Proof: Eqn. (12) and (13) in Lemma 1 are obtained from the constraints (10) and (11) directly by substituting the corresponding utility and payment functions. Constraint (12) is proved to hold with equality in equilibrium [15], which gives Eqn. (14) in Lemma 1.

Considering the Lagrangian function of cloud provider 1's problem, the requirement that the Lagrangian multipliers corresponding to inequality constraints should be greater or equal to 0 gives Eqn. (15) and (17) in Lemma 1. Similarly, by considering the Lagrangian function of the cloud provider 2's problem, we have Eqn. (16) and (18). As a result, all equations in Lemma 1 are necessary for the existence of a Nash equilibrium. ∎

Having results from Lemma 1, we now come to the sufficient condition for the Nash equilibrium, which is stated in Theorem 1. Due to the complexity of this problem, we are not able to obtain the exact analytical presentation of the price and the market share of each cloud provider in equilibrium. However, we have obtained an important relation between equilibrium prices, which shows the pricing gap between two cloud providers.

Theorem 1: Let $(p_1^*, p_2^*, f_1^*, f_2^*)$ be a feasible solution that satisfies all equations in Lemma 1. Then (p_1^*, p_2^*) is a Nash equilibrium if it satisfies:

$$p_1^* - p_2^* = \frac{c(\mu_1 - \mu_2)}{\mu_1\mu_2} - \frac{cf_1^*}{\Lambda\mu_1(\mu_1 - f_1^*)} + \frac{cf_2^*}{\Lambda\mu_2(\mu_2 - f_2^*)}$$

and

$$h'(\hat{f}_i) \leq 0,$$

where

$$\begin{aligned} h(f_i) &= \Bigg[p_k + \frac{c(\mu_i - \mu_k)}{\mu_i\mu_k} - \frac{cf_i}{\Lambda\mu_i(\mu_i - f_i)} \\ &+ \frac{c(\Lambda - f_i)}{\Lambda\mu_k(\mu_k - \Lambda + f_i)}\Bigg] f_i - \gamma_i(\mu_i), \end{aligned}$$

$i \neq k$ and $i, k = 1, 2$. $\hat{f}_i$ is obtained by solving

$$h''(f_i) = 0.$$

Proof: We take the cloud provider 1's problem as an example to show a proof sketch. In view of Lemma 1, cloud provider 1's problem is equivalent to

$$\begin{aligned} \max_{0 \leq f_1 \leq \Lambda} \quad & h(f_1) = \Bigg[p_2 + \frac{c(\mu_1 - \mu_2)}{\mu_1\mu_2} - \frac{cf_1}{\mu_1\Lambda(\mu_1 - f_1)} \\ & + \frac{c(\Lambda - f_1)}{\mu_2\Lambda(\mu_2 - \Lambda + f_1)}\Bigg] f_1 - \gamma_1(\mu_1) \\ \text{s.t.} \quad & \frac{Mv}{f_1} + \frac{Mc\Lambda - cf_1}{\mu_1\Lambda(\mu_1 - f_1)} \\ & + \frac{c(\Lambda - f_1)}{\mu_2\Lambda(\mu_2 - \Lambda + f_1)} \geq r - p_2 - \frac{c}{\mu_2}, \end{aligned}$$

with additional constraints on f_1 such that $f_1 < \mu_1$ and $f_2 = \Lambda - f_1 < \mu_2$. By differentiation, we show that function h' is concave. Let $\hat{f}_1$ be the maximal point of h' in $(\Lambda - \mu_2, \mu_1)$, $\hat{f}_1$ is the root of

$$h''(f_1) = 0. \tag{19}$$

To guarantee Inequality (12) to be binding in the case of two cloud providers, we have to further require $h'(\hat{f}_1) \leq 0$, which completes the proof. ∎

Corollary 1: When $\mu_2 > \mu_1$, $p_2^* > p_1^*$ in equilibrium.

Since cloud users are sensitive to finishing times of their tasks, a cloud provider with a larger resource capacity is more likely to complete a task within a shorter period of time. As a consequence, it is preferred by more cloud users. Because a larger resource capacity helps a cloud provider to enjoy an advantageous position in the competition market, the cloud provider can charge a higher price as a result.

Proof: Given $\mu_2 > \mu_1$ and $\mu_1 > f_1^*$

$$\begin{aligned} p_2^* - p_1^* &= \frac{c(\mu_2 - \mu_1)}{\mu_1\mu_2} + \frac{cf_1^*}{\Lambda\mu_1(\mu_1 - f_1^*)} - \frac{cf_2^*}{\Lambda\mu_2(\mu_2 - f_2^*)} \\ &> \frac{f_1^*\mu_2^2 - \mu_2 f_2^* f_1^* - f_2^*\mu_1^2 + \mu_1 f_2^* f_1^*}{\mu_1\mu_2(\mu_1 - f_1^*)(\mu_2 - f_2^*)} \\ &> \frac{f_1^*\mu_2(\mu_2 - f_2^*)}{\mu_1\mu_2(\mu_1 - f_1^*)(\mu_2 - f_2^*)} > 0. \end{aligned}$$

∎

Based on the results in Theorem 1, we present a simple iterative algorithm that can be used to compute the price for each cloud provider in a duopoly market, described in Algorithm 1. Though we are unable to prove that the converged price is guaranteed to be the Nash equilibrium, our simulation results shown in Sec. V have shown that this is the case in our simulation scenarios.

Algorithm 1 *Compute the price for cloud provider i in a duopoly market.*

1: (*Initialization*). Each cloud provider i sets the usage price to be $p_i = r - v - c$.
2: (*Iterative step*). Each cloud provider i then updates its price using another cloud provider i's price and the current market shares f_i and f_i':
$p_i = p_i' + \frac{c(\mu_i - \mu_i')}{\mu_i\mu_i'} - \frac{cf_i}{\Lambda\mu_i(\mu_i - f_i)} + \frac{cf_i'}{\Lambda\mu_i'(\mu_i' - f_i')}$
3: (*Convergence criterion*). Repeat the *iterative step* until the price p_i differs from its previous value by less than a predetermined value ϵ.

We use an example to illustrate how prices converge iteratively. Since the operating cost of each cloud provider does not affect their selections of prices, we set both $\gamma_1(\mu_1)$ and $\gamma_2(\mu_2)$ to 0 in this example. Suppose the reservation value $v = 1$, the benefit factor $r = 5$, and the waiting cost factor $c = 1$. When there are 20 cloud users in the market, the convergence of usage prices per resource unit as well as the market shares of two cloud providers with resource capacities $\mu_1 = 2$ and $\mu_2 = 4$ are shown in Fig. 2 and Fig. 3, respectively. We set $\epsilon = 0.001$ as the convergence criterion in this example.

As we can see, the proposed algorithm converges rapidly — within 4 iterations. As we have shown in Corollary 1, the cloud provider with a larger resource capacity charges a higher price. In this case, $p_2^* > p_1^*$. We can also observe from the

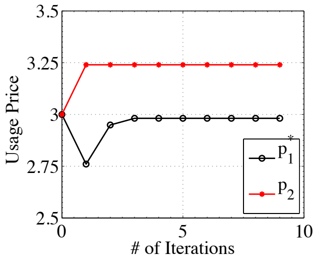


Fig. 2. The convergence of usage prices.

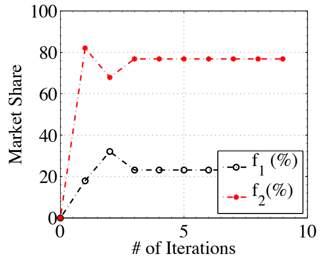


Fig. 3. The convergence of market shares.

figure that the cloud provider with a larger resource capacity also attracts more cloud users in equilibrium. This reveals that cloud providers can become more competitive in the market by increasing its resource capacity. Since cloud users are sensitive to finishing times of their tasks, even a cloud provider with a higher resource capacity will charge a higher price, most cloud users still prefer to choose this faster provider.

2) Nash Equilibrium in a Duopoly Market with Homogeneous Cloud Providers: In the homogeneous case that two cloud providers have the same resource capacity, the next theorem establishes the result that a unique Nash equilibrium exists in the duopoly market, and that the iterative algorithm above always converges to the same price for both cloud providers. It can be derived by solving the optimization problem in Theorem 1.

***Theorem** 2:* When reduced to homogeneous cloud providers, *i.e.*, $\mu_1 = \mu_2 = \mu$, the Nash equilibrium (p_1^*, p_2^*) in Theorem 1 takes the form of:

$$p_1^* = p_2^* = r - \frac{c}{\mu} - \frac{2Mc}{\mu(2\mu - \Lambda)} - \frac{2Mv}{\Lambda}$$

with the corresponding market shares $f_1^* = f_2^* = \Lambda/2$.

Proof: When the two cloud providers are equivalent to each other in service rate, *i.e.*, $\mu_1 = \mu_2 = \mu$, the two cloud providers are indifferent to the cloud users, which implies that the equilibrium solution is symmetric [26]. According to Theorem 1, we have

$$p_1^* - p_2^* = \frac{c(f_2^* - f_1^*)}{\Lambda(\mu - f_1^*)(\mu - f_2^*)}. \quad (20)$$

Based on constraint (11), we know that the following relationship must hold between the optimal prices and market shares.

$$f_1^* + f_2^* = \frac{c(f_1^* - f_2^*)}{(p_1^* - p_2^*)(\mu - f_1^*)\mu - f_2^*}. \quad (21)$$

Substitute Eqn. (20) into Eqn. (21), we can obtain that

$$f_1^* + f_2^* = \Lambda.$$

Based on the symmetric characteristic of the equilibrium solution, we now have that

$$f_1^* = f_2^* = \Lambda/2.$$

Now let's take cloud provider 1's expected profit maximization problem as an example. Constraint (10) requires that

$$(r\mu - p_1\mu - c)(\mu - f_1)f_1 \geq Mv\mu(\mu - f_1) + Mcf_1.$$

When $f_1 = \Lambda/2$, this implies that

$$p_1 \leq r - \frac{c}{\mu} - \frac{2Mc}{\mu(2\mu - \Lambda)} - \frac{2Mv}{\Lambda}.$$

Since the optimal price is the one that maximizes the expected profit of each cloud provider, which equals $p_i f_i$. We get he result that the Nash equilibrium price $p_1^* = r - \frac{c}{\mu} - \frac{2Mc}{\mu(2\mu-\Lambda)} - \frac{2Mv}{\Lambda}$. Similarly, we can obtain that the Nash equilibrium price $p_2^* = r - \frac{c}{\mu} - \frac{2Mc}{\mu(2\mu-\Lambda)} - \frac{2Mv}{\Lambda}$ as well, which completes the proof. ■

We can see from Theorem 2 that, when cloud providers are homogeneous, *i.e.*, they have the same resource capacities, both cloud providers will charge the same price, which has the form of the monopoly price p_Λ, with each of them taking half of the market. That is, both cloud providers behave independently and operate exactly the same as monopolists.

***Corollary** 2:* The comparative statics of the homogeneous Nash equilibrium price in Theorem 2 are as follows:

$$\begin{aligned}
\frac{\partial p^*}{\partial r} &= 1 > 0, \ \frac{\partial p^*}{\partial c} = -\frac{1}{2}\sqrt{\frac{1}{\mu c}} < 0 \\
\frac{\partial p^*}{\partial v} &= -\frac{M}{\Lambda} < 0, \ \frac{\partial p^*}{\partial \mu} = \frac{1}{2\mu}\sqrt{\frac{c}{\mu}} + \frac{Mc(2\mu - \Lambda)}{\mu_2(\mu - \Lambda)^2} > 0 \\
\frac{\partial p^*}{\partial \Lambda} &= \frac{2M}{\mu\Lambda^2(2\mu - \Lambda)^2}\left[(\mu v - c)\Lambda^2 - 4v\mu^2\Lambda + 4v\mu^3\right] \\
&\begin{cases} \begin{cases} > 0 & 0 < \Lambda < \frac{2v\mu^2 - 2\mu\sqrt{\mu v c}}{\mu v - c} \\ < 0 & \frac{2v\mu^2 - 2\mu\sqrt{\mu v c}}{\mu v - c} < \Lambda < \mu \end{cases} & \mu > \frac{c}{v} \\ > 0 \quad 0 < \Lambda < \mu & \mu \leq \frac{c}{v} \end{cases}
\end{aligned}$$

The comparative statics of the equilibrium price are illustrated in Fig. 4. As we can see, they conform with most of our intuitions. A cloud provider will raise the usage price with an increase in the users' benefit factor r and its resource capacity μ, and reduce the price in response to an increase in the waiting cost factor c and the reservation value v.

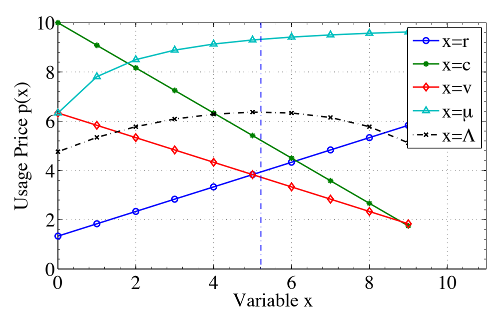


Fig. 4. An illustration of the statics of the homogeneous Nash equilibrium price.

Since the benefit factor r reflects the importance of the task and the waiting cost factor c represents its urgency, this implies that the more important the task is, the more the cloud provider will charge the user; the more urgent users view the task, the less the cloud provider "dares" to charge in order to win the "deal" from users. The rationale is that if the cloud provider knows that the task is important to the user, *i.e.*, the user will gain substantial benefit by completing the task, the cloud provider will infer that the user is willing to pay more to finish the task, and hence ask for more. Yet if the task is urgent, the cloud provider will tend to ask less to make up for its limited resource capacity.

In addition, if the cloud provider is able to increase its resource capacity μ by investing in new servers or by upgrading its current facilities, it will be able to ask more for the improved service quality. In the figure, we can see that when the resource capacity is small, the demand will be relatively strong compared to the resource capacity, which results in a non-competitive environment in the market. Increasing the resource capacity slightly results in a rapid increase of usage price. Finally, in cases when users have a higher reservation value v and thus may refuse to use the cloud service with a higher probability, cloud providers will have to reduce their prices to attract users.

The market size Λ will affect the usage price in a more complicated way. When the resource capacity is large enough, *i.e.*, $\mu > \frac{c}{v}$, the cloud provider will raise the price when the market size increases, until it reaches a certain threshold, *e.g.*, $\Lambda = 4.6$ in the figure. If the market size continues to increase, users will overwhelm the cloud providers and may experience longer task finishing times. Cloud providers will reduce the price to compensate for the increased waiting costs. However, if the resource capacity is small, *i.e.*, $\mu < \frac{c}{v}$, the market is again in a non-competitive environment, which results in the fact that price increases with a larger market size.

3) Social Welfare Problem in the Duopoly Market: We have previously analyzed the Nash equilibrium prices in a competition market with two cloud providers. In equilibrium, each cloud provider's price is determined by its best response function to the other cloud provider's price. In other words, prices are optimized for cloud providers only, with no direct implication that all cloud providers and users will reach an outcome that is *socially optimal*. A choice of prices — one by each cloud provider — is socially optimal if it maximizes the sum of payoffs to all participants [27]. In the cloud market, it implies a set of equilibrium prices at which the payoffs of both cloud providers and cloud users are maximized. A cloud user j's payoff for being served by cloud provider i is its expected net rewards $U(\lambda_j) - P_{ij}$, with a request rate of λ_j and a usage price p_i, $i = 1, 2$; a cloud provider i's payoff in this market equals its expected profit, which is π_i. Therefore, the social welfare is:

$$\sum_{i,j} [U(\lambda_j) - P_{ij}] + \sum_i \pi_i.$$

In the social welfare problem, prices are simply an internal transfer of wealth and hence are not considered as objective variables. Our interest is how cloud users are distributed between two cloud providers to maximize social welfare. Though we hope that the duopoly equilibrium prices are also socially optimal, our analysis shows it is not always the case.

***Theorem** 3:* The social welfare maximizing solution (f_{1s}^*, f_{2s}^*) is not always the same as the market shares (f_1^*, f_2^*) in equilibrium.

Proof: The social welfare maximization problem in a duopoly cloud market takes the form of:

$$\begin{aligned}
\max_{f_1<\mu_1, f_2<\mu_2} \quad & (r - \frac{c}{\mu_1})f_1 - \frac{cMf_1}{\mu_1(\mu_1 - f_1)} + (r - \frac{c}{\mu_2})f_2 \\
& - \frac{cMf_2}{\mu_2(\mu_2 - f_2)} - 2Mv - \gamma_1(\mu_1) - \gamma_2(\mu_2) \\
\text{s.t.} \quad & f_1 + f_2 \le \Lambda
\end{aligned}$$

The first-order condition for the social welfare problem is

$$\begin{aligned}
& r - \frac{c}{\mu_i} - \frac{cM}{(\mu_i - f_i)^2} - \eta = 0 \ , \ i = 1, 2, \\
& \eta(\Lambda - f_1 - f_2) = 0,
\end{aligned}$$

where η is the Lagrangian multiplier. When $f_1 + f_2 < \Lambda$, we get the social optimal solution equals:

$$f_{is}^* = \mu_i - \sqrt{\frac{cM\mu_i}{r\mu_i - c}} \ , \ i = 1, 2.$$

Otherwise when $f_1 + f_2 = \Lambda$, the social optimal solution is given by

$$f_{is}^* = \mu_i - \sqrt{\frac{cM\mu_i}{r\mu_i - c - \eta\mu_i}} \ , \ i = 1, 2,$$

where η is the unique solution to

$$\mu_1 - \sqrt{\frac{cM\mu_1}{r\mu_1 - c - \eta\mu_1}} + \mu_2 - \sqrt{\frac{cM\mu_2}{r\mu_2 - c - \eta\mu_2}} = \Lambda \quad (22)$$

in the interval $(0, \min\{r - \frac{c}{\mu_1}, r - \frac{c}{\mu_2}\})$. The reason that η is uniquely determined by Eqn. (22) in a certain interval is that the left-hand-side of Eqn. (22) is monotonically decreasing in η. The value of the left-hand-side expression is greater than Λ when $\eta = 0$, and is approaching $-\infty$ as η increases to $\min\{r - \frac{c}{\mu_1}, r - \frac{c}{\mu_2}\}$.

It is not difficult to verify that the market shares in equilibrium are socially optimal in a homogeneous duopoly cloud market, *i.e.*, when two cloud providers have the same service rate $\mu_1 = \mu_2 = \mu$. Since we have shown that $f_1 + f_2 = \Lambda$ in this case in Theorem 2, the social optimal market share of both cloud providers takes the form

$$f_{1s}^* = f_{2s}^* = \mu - \sqrt{\frac{cM\mu}{r\mu - c - \eta\mu}},$$

where η can be obtained from the equation

$$2\mu - 2\sqrt{\frac{cM\mu}{r\mu - c - \eta\mu}} = \Lambda.$$

Combine these two equations together, we can get the result that the social optimal market share in a homogeneous duopoly cloud market is $f_{1s}^* = f_{2s}^* = \Lambda/2$. We can see that the social optimal market share equals to the equilibrium market share $f_1^* = f_2^* = \Lambda/2$ that we have proved in Theorem 2, which leads to a Price of Anarchy (PoA) of 0 in this case. However, we can also see that in a more general case when $f_1 + f_2 < \Lambda$, the Nash equilibrium is not typically socially optimal. ■

Though the conclusion that the social welfare maximizing solution is not the same as the market shares in equilibrium is not a surprise, it is not intuitive either. More importantly, a Price of Anarchy of 0 can be achieved in a homogeneous duopoly, which means that the social welfare maximizer also reflects the equilibrium market share.

B. The General Case

Based on our game theoretical analysis of price competition in the monopoly and duopoly cloud markets, we now proceed to consider the general case when multiple cloud providers are competing with one another. Our analyses will show that a unique Nash equilibrium exists in an oligopoly cloud market. We will also present an iterative algorithm to compute the equilibrium prices based on our analyses.

1) Cloud Provider i's Problem in an Oligopoly Market: From our previous analysis in the duopoly market, we can see that the market share of each cloud provider is not only affected by the cloud provider's own price, but also the other cloud provider's pricing choice, which are both variables to be determined. In a market with multiple cloud providers, the usage prices will influence each cloud provider's market share in a highly complicated way, and due to this reason we are not able to get the exact analytical presentation. As an alternative, We apply the Multiplicative Competitive Interaction (MCI) model to capture the relationship between usage prices and market shares in an oligopoly market. Proposed by Bell *et al.* in 1975 [28], the MCI model is widely used in economic competition markets [16], [26].

To be specific, the market share of each cloud provider in a market with N cloud providers is assumed to take the following form:

$$f_i = \Lambda \left(\frac{L_i p_i^{-a} t_i^{-b}}{\sum_{j=1}^{N} L_j p_j^{-a} t_i^{-b}} \right), \forall i = 1, 2, ..., N,$$

where t_i represents the expected finishing time of a unit request experienced at cloud provider i, including both the waiting time in the queue and the service time. The numerator $L_i p_i^{-a} t_i^{-b}$, with $a, b \geq 0$, represents the *attraction* of cloud provider i, which corresponds to how cloud users feel towards its service given its usage price, expected finishing time, and other competitive factors, *e.g.*, API, load balancing, and reputation. The parameter a and b are referred to as the price attraction factor and the finishing time attraction factor, respectively. The parameter $L_i > 0$ represents the combined effects of other competitive factors, with a larger L_i reflecting a higher degree of attraction to cloud users. Cloud provider i's market share is given by its relative attraction to all cloud providers in the market. In subsequent analyses, we choose $a = b = 1$ for simplicity.

Based on queueing theory, the expected finishing time t of a single request, including both the waiting time and the service time, equals $\frac{1}{\mu - f}$ in an M/M/1 queue. Note that in a cloud environment, for a given resource capacity μ_i at cloud provider i, the expected finishing time is a function of its market share f_i, which is determined by the cloud provider's usage price. As a consequence, if we use $\beta_i = \frac{1}{L_i} \sum_{j \neq i} L_j p_j^{-1} t_j^{-1}$ to denote the combined attraction of cloud provider i's competitors, the market share of cloud provider i is

$$f_i = \frac{\Lambda}{\beta_i p_i t_i + 1}, \forall i = 1, 2, ..., N.$$

Substituting $t_i = \frac{1}{\mu_i - f_i}$, together with the requirement that $f_i < \mu_i$, we have

$$t_i = \frac{2}{\mu_i - \beta_i p_i - \Lambda + \sqrt{(\mu_i + \beta_i p_i + \Lambda)^2 - 4\Lambda\mu_i}}. \quad (23)$$

Expressed as a function of usage prices, the market share of cloud provider i, f_i, is in a much more complicated form than that in typical economic papers in the literature discussing price competition, and this makes our subsequent analyses substantially more challenging.

Again, the objective of cloud provider i is to find the best response function that maximize its expected profit, taking into consideration the usage prices set by other cloud providers. Mathematically, the problem faced by cloud provider i can be formulated as the following:

$$\begin{aligned}
\max_{p_i \geq 0} \quad & \pi_i = \Lambda \left(\frac{p_i}{\beta_i p_i t_i + 1} \right) - \gamma_i(\mu_i) & (24) \\
\text{s.t.} \quad & \frac{\Lambda(\mu_i r - c - \mu_i p_i)}{\mu_i(\beta_i p_i t_i + 1)} - \frac{cM\Lambda}{\mu_i(\mu_i \beta_i p_i t_i + \mu_i - \Lambda)} \geq Mv \\
& \mu_i > \frac{\Lambda}{\beta_i p_i t_i + 1} \geq 0 \\
& t_i = \frac{2}{\mu_i - \beta_i p_i - \Lambda + \sqrt{(\mu_i + \beta_i p_i + \Lambda)^2 - 4\Lambda\mu_i}}.
\end{aligned}$$

Constraints in problem (24) ensure that any cloud user who chooses cloud provider i will be offered at least the same expected net reward as its reservation value.

2) Nash Equilibrium in an Oligopoly Market: In an oligopoly market with N cloud providers, the N-tuple price vector $(p_1^*, p_2^*, ..., p_N^*)$ is called a Nash equilibrium if for each cloud provider i, p_i^* is the best response to price p_j^* chosen by all other firms $j \neq i$. In other words, the Nash equilibrium implies that no single cloud provider can benefit by deviating from this equilibrium point unilaterally. By solving problem (24), cloud provider i is able to compute its optimal price p_i based on the combined attraction of its competitors β_i, which in turn can be computed using the current prices p_j of other cloud providers $j \neq i$. With the idea of solving this problem in an iterative fashion, we have designed the following iterative algorithm, Algorithm 2, to compute the Nash equilibrium price for each cloud provider.

In subsequent analyses, we are going to show that a unique Nash equilibrium exists in a price competition market with multiple cloud providers, and the above iterative algorithm always converges to this equilibrium solution. This is fairly significant, in that if the required information in Algorithm 2 is available, we now have an algorithmic tool to compute the unique Nash equilibrium. We first present a necessary result that is useful to derive the key results in Lemma 3.

Lemma 2: When $\mu_i > \Lambda$, $Mv\mu_i^2 + cM\Lambda > \Lambda^2(\mu_i r - c)$ and $Mv\mu_i(\mu_i - \Lambda) + cM\Lambda < \Lambda(\mu_i r - c)(\mu_i - \Lambda)$.

Proof: We use the first inequality to show a proof sketch here. The second inequality can be proved by using the same

Algorithm 2 *Compute the Nash equilibrium price for cloud provider i in an oligopoly market.*

1: (*Initialization*). Each cloud provider i sets the usage price to be a very small value $p_i = \epsilon'$.
2: (*Iterative step*).
3: **for** cloud provider 1 to N **do**
4: Each cloud provider i computes the optimal price p_i by solving problem (24) using the current values p_j of other cloud providers $j \neq i$, and updates the price p_i.
5: **end for**
6: (*Convergence criterion*). Repeat the *iterative step* until price p_i differs from its previous value by less than some predetermined value ϵ.

technique. The first inequality in Lemma 2 is equivalent to

$$Mv\mu_i^2 - \Lambda^2 r\mu_i + \Lambda^2 c - cM\Lambda > 0. \quad (25)$$

It is obvious that $Mv > 0$, $-\Lambda^2 r < 0$. Since $\Lambda \geq M$, $\Lambda^2 c - cM\Lambda \geq 0$. If $\Lambda^4 r^2 - 4Mv(\Lambda^2 c - cM\Lambda) \leq 0$, then inequality (25) holds for all μ_i. On the other hand, if $\Lambda^4 r^2 - 4Mv(\Lambda^2 c - cM\Lambda) > 0$, we find that $\Lambda > \overline{\mu_i}$, where $\overline{\mu_i}$ is the larger root given by the quadratic function in inequality (25). To summarize, inequality (25) holds when $\mu_i > \Lambda$, which completes the proof. ■

Based on Lemma 2, the following Lemma establishes two key results for analyzing the price competition in an oligopoly cloud market. To be specific, Lemma 3 characterizes the optimal price for each cloud provider when prices of all other providers are given. It also reveals the fact that a cloud provider will have to compete with a lower price when the combined attraction of its competitors increases. Furthermore, Lemma 3 also serves as an important supporting lemma in the proof of Theorem 4.

***Lemma** 3:* (i) A unique optimal price p_i^* exists for cloud provider i, and is bounded by $0 < p_i^* \leq \overline{p_i}$. (ii) The optimal price p_i^* is decreasing in β_i.

Proof: According to Eqn. (23), the expected finishing time t_i is a function of the usage price p_i at each cloud provider i. Let $H(p_i) = p_i t_i$. $H(p_i)$ takes the form of

$$H(p_i) = \frac{2p_i}{\mu_i - \beta_i p_i - \Lambda + \sqrt{(\mu_i + \beta_i p_i + \Lambda)^2 - 4\Lambda\mu_i}}.$$

Note that β_i is determined by the price and combined effects of all other cloud provides except i. Once the prices of other cloud providers are fixed, β_i is a constant from cloud provider i's point of view. Since

$$\frac{\partial H(p_i)}{\partial p_i} = \frac{2(\alpha_i + \sqrt{\alpha_i'^2 - 4\Lambda\mu_i}) + 2p_i \frac{\beta_i\alpha_i' - \beta_i\sqrt{\alpha_i'^2 - 4\Lambda\mu_i}}{\sqrt{\alpha_i'^2 - 4\Lambda\mu_i}}}{\alpha_i + \sqrt{\alpha_i'^2 - 4\Lambda\mu_i}},$$

where $\alpha_i = \mu_i - \beta_i p_i - \Lambda$, $\alpha_i' = \mu_i + \beta_i p_i + \Lambda$, we know that $\frac{\partial H(p_i)}{\partial p_i} > 0$, which implies that $H(p_i)$ is increasing with p_i. Fig. 5 shows how $H(p_i)$ changes with price p_i. As we can see in the figure, if we can prove that a uniquely existed $H^*(p_i)$ that maximizes cloud provider i's expected profit in problem (24) is bounded by $[0, \overline{H(p_i)}]$, and is decreasing in β_i, we are able to prove that a uniquely existed optimal price p_i^* for cloud provider i is bounded by $0 < p_i^* \leq \overline{p_i}$, and is decreasing in β_i.

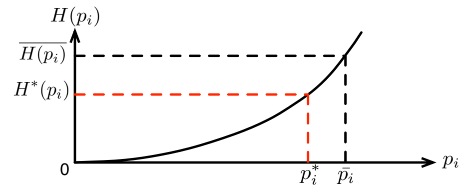


Fig. 5. The graph of the function $H(p_i)$.

(i) Based on the fact that the equilibrium market share $f_i < \mu_i$, we have

$$H(p_i) \geq \frac{\Lambda - \mu_i}{\beta_i\mu_i}.$$

Since $H(p_i) = p_i t_i$ should also be greater than 0 to ensure a non-negative expected profit obtained by cloud provider i, the inequality

$$\frac{\Lambda - \mu_i}{\beta_i\mu_i} < 0 \Rightarrow \Lambda < \mu_i \quad (26)$$

must hold to avoid an infinite queueing delay.

The constraints in problem (24) can be rephrased as

$$A_i H(p_i)^2 + B_i H(p_i) + C_i \leq 0, \quad (27)$$

where $A_i = Mv\mu_i^2\beta_i^2 + \mu_i^2\Lambda\beta_i$, $B_i = Mv\mu_i\beta_i(\mu_i - \Lambda) + \mu_i^2\beta_i Mv - \mu_i\beta_i\Lambda(\mu_i r - c) + \mu_i\Lambda(\mu_i - \Lambda) + cM\Lambda\beta_i$, and $C_i = Mv\mu_i(\mu_i - \Lambda) + cM\Lambda - \Lambda(\mu_i r - c)(\mu_i - \Lambda)$. It is obvious that $A_i > 0$. For a meaningful problem that is worth discussing, each cloud provider should be able to attract at least one cloud user when it charges a usage price 0 and there is no other cloud users waiting in the queue, which is to say

$$\left(r - \frac{c}{\mu_i}\right)\Lambda > Mv \Rightarrow (r\mu_i - c)\Lambda > Mv\mu_i. \quad (28)$$

Based on Eqn. (28) and the result in Lemma 2, we have $C_i < 0$ and $B_i \geq \mu_i\Lambda(\mu_i - \Lambda)$, $\forall i$.

This shows that the quadratic function in (27) has two real-value roots $\underline{H(p_i)}$ and $\overline{H(p_i)}$, with $\underline{H(p_i)} < 0$ and $\overline{H(p_i)} > 0$. This establishes the result that $H^*(p_i)$ is bounded by $0 < H^*(p_i) \leq \overline{H(p_i)}$, which implies that the equilibrium price p_i^* for cloud provider i is bounded.

(ii) According to Eqn. (27), we have

$$(2Mv\mu_i^2\beta_i + \mu_i^2\Lambda)H(p_i)^2 + (Mv\mu_i^2\beta_i^2 + \mu_i^2\beta_i\Lambda)2H(p_i) \times \frac{\partial H(p_i)}{\partial \beta_i} + \frac{1}{\beta_i}[B_i - \mu_i\Lambda(\mu_i - \Lambda)]H(p_i) + B_i\frac{\partial p_i}{\partial \beta_i} \leq 0,$$

which implies that

$$\frac{\partial H(p_i)}{\partial \beta_i} \leq \frac{-H(p_i)^2 - \frac{H(p_i)}{\beta_i}\frac{B_i - \mu_i\Lambda(\mu_i - \Lambda)}{2Mv\mu_i^2\beta_i + \mu_i^2\Lambda}}{2H(p_i)\frac{Mv\mu_i^2\beta_i^2 + \mu_i^2\beta_i\Lambda}{2Mv\mu_i^2\beta_i + \mu_i^2\Lambda} + B_i} \leq 0.$$

This shows that function $H(p_i)$ is decreasing in β_i. Based on the fact that

$$\frac{\partial p_i}{\partial H(p_i)} = \frac{2(\mu_i - \Lambda)}{H(p_i)^3[\sqrt{\frac{2}{H(p_i)^2} - \beta_i^2}]^{\frac{3}{2}}} \geq 0,$$

we have $\frac{\partial p_i}{\partial \beta_i} = \frac{\partial p_i}{\partial H(p_i)} \frac{\partial H(p_i)}{\partial \beta_i} \leq 0$, which establishes the second statement that the optimal price p_i^* is decreasing in β_i. We further observe that

$$\frac{\partial \pi_i}{\partial p_i} = \frac{\Lambda - \Lambda p_i^2 t_i \frac{\partial \beta_i}{\partial p_i} - \Lambda^2 p_i^2 \beta_i \frac{\partial t_i}{\partial p_i}}{(\beta_i p_i t_i + 1)^2}.$$

Based on the fact that $\frac{\partial \beta_i}{\partial p_i} < 0$ and $\frac{\partial t_i}{\partial p_i} < 0$, we have $\frac{\partial \pi_i}{\partial p_i} > 0$, which completes the statement that the optimal price p_i^* uniquely exists.

■

Theorem 4: Algorithm 2 always converges to the unique Nash equilibrium solution in an oligopoly cloud market.

Proof: We are now ready to show that Algorithm 2 always converges to the Nash equilibrium. Let $p_i^{(k)}$ be the optimal price at the kth iteration. We shall prove by induction that $p_i^{(k)}$ is increasing in k. Since p_i^* is bounded by $\overline{p_i}$, this establishes the result that $p_i^{(k)}$ always converges.

By definition, $p_i^{(0)} = \epsilon'$, where ϵ' can be chosen to be small enough such that $p_i^{(k)} > \epsilon'$. In particular, when $k = 1$, $p_i^{(1)} > p_i^{(0)}$. Now assume that $p_i^{(k)} \geq p_i^{(k-1)}$ for all i and $k < n$. At the beginning of the nth iteration, we have

$$\begin{aligned} \beta_1^{(n)} &= \frac{1}{L_1} \sum_{j>1} L_j (p_j^{(n-1)})^{-1} \\ &\leq \frac{1}{L_1} \sum_{j>1} L_j (p_j^{(n-2)})^{-1} = \beta_1^{(n-1)}, \end{aligned}$$

where the inequality follows from the inductive assumption.

Then from the second statement in Lemma 3, we know that $p_1^{(n)} \geq p_1^{(n-1)}$. Suppose that $p_j^{(n)} \geq p_j^{(n-1)}$, for all $j = 1, 2, ..., l-1$. Then for cloud provider l at the nth iteration we have

$$\begin{aligned} \beta_l^{(n)} &= \frac{1}{L_l} \sum_{j>l} L_j (p_j^{(n)})^{-1} + \frac{1}{L_l} \sum_{j>l} L_j (p_j^{(n-1)})^{-1} \\ &\leq \frac{1}{L_l} \sum_{j>l} L_j (p_j^{(n-1)})^{-1} + \frac{1}{L_l} \sum_{j>l} L_j (p_j^{(n-2)})^{-1} \\ &= \beta_l^{(n-1)}, \end{aligned}$$

As a result, we deduct that $p_l^{(n)} \geq p_l^{(n-1)}$. This implies that $p_i^{(n)} \geq p_i^{(n-1)}$ for all i, which completes our proof of convergence.

We now prove that the converged Nash equilibrium must be unique by contradiction. The equilibrium result can be represented as $\Phi^* = (At_1^*, At_2^*, ..., At_N^*)$, where $At_i^* = L_i p_i^{-1*} t_i^{-1*}$ represents the attraction of cloud provider i given the converged equilibrium price p_i^*. Since we have proved in Lemma 3 that the optimal price p_i^* for each cloud provider uniquely exists, the corresponding expected finishing time of a unit request experienced at cloud provider i t_i^* is also uniquely determined by Eqn. (23), which uniquely defines each At_i^* in Φ^*.

Let us assume that there exists another equilibrium result that can be represented as $\Phi' = (At_1', At_2', ..., At_N')$, where each At_i' is determined by the price p_i'. Express $At_i' = (1 + \theta_i)At_1^*$. By numbering the cloud providers properly, we can assume that $\theta_1 \geq \theta_2 \geq ... \geq \theta_N$ and $\theta_1 > 0$.

Our first argument is that $\theta_2 > 0$. If $\theta_2 \leq 0$, it will imply that $\theta_i \leq 0, \forall i \geq 2$, which is equivalent to say that $At_i' \leq At_i^*, \forall i \geq 2$. Since for cloud provider 1, the combined attraction of all its competitors $\beta_1 = \frac{1}{L_1} \sum_{j \geq 2} At_j$, we have $\beta_1' \leq \beta_1^*$. Recall in our proof of Lemma 3, we proved that the function $H(p_i) = p_i t_i$ is decreasing in β_i, which implies that $p_1' t_1' \geq p_1^* t_1^*$. We can then express the equilibrium attraction of cloud provider 1 as

$$At_1^* = L_1 p_1^{*-1} t_1^{*-1} \geq p_i^{'-1} t_i^{'-1} = At_1' = (1 + \theta_1)At_1^*. \quad (29)$$

Eqn. (29) implies that $\theta_1 \leq 0$, which contradicts with our assumption that $\theta_1 > 0$.

Now, let us consider the optimization problem of cloud provider 1 with the attractions of other cloud providers to be $At_i'' = (1+\theta_2)At_i^*, \forall i \geq 2$. Let p_1'' and t_1'' be the corresponding optimal solution solved by cloud provider 1, then from the first constraint in problem (24), we have

$$\frac{\Lambda(\mu_1 r - c - \mu_1 p_1'')}{\mu_1(\beta_1'' p_1'' t_1'' + 1)} - \frac{cM\Lambda}{\mu_1(\mu_1 \beta_1'' p_1'' t_1'' + \mu_1 - \Lambda)} = Mv \quad (30)$$

Since $\Phi^* = (At_1^*, At_2^*, ..., At_N^*)$ is an equilibrium result, we also have

$$\frac{\Lambda(\mu_1 r - c - \mu_1 p_1^*)}{\mu_1(\beta_1^* p_1^* t_1^* + 1)} - \frac{cM\Lambda}{\mu_1(\mu_1 \beta_1^* p_1^* t_1^* + \mu_1 - \Lambda)} = Mv \quad (31)$$

Since we have proved that $\theta_2 > 0$ and $At_i'' = (1 + \theta_2)At_i^*, \forall i \geq 2$, we have $\beta_1'' > \beta_1^*$. Again, we can represent the functions in Eqn. (30) and Eqn. (31) as $A_1'' H(p_1'')^2 + B_1'' H(p_1'') + C_1'' = 0$ and $A_1^* H(p_1^*)^2 + B_1^* H(p_1^*) + C_1^* = 0$, where A_1'', B_1'', C_1'' and A_1^*, B_1^*, C_1^* take the same form as A_i, B_i, C_i in Eqn. (27). Since $\beta_1'' > \beta_1^*$, we have $A_1'' > A_1^*$. Since $\frac{\partial B_1}{\beta_1} = \frac{B_1 - \mu_1 \Lambda(\mu_1 - \Lambda)}{\beta_1}$, and we have $B_1 \geq \mu_1 \Lambda(\mu_1 - \Lambda)$, this implies that $B_1'' > B_1^*$. Based on the fact that $\frac{B_1}{A_1} = \frac{Mv\mu_1(\mu_1-\Lambda)+\mu_1^2 Mv - \mu_1 \Lambda(\mu_1 r - c) + cM\Lambda}{Mv\mu_1^2 \beta_1 + \mu_1^2 \Lambda} + \frac{\mu_1 \Lambda(\mu_1 - \Lambda)}{Mv\mu_1^2 \beta_1^2 + \mu_1^2 \Lambda \beta_1}$, we have $\frac{B_1''}{A_1''} < \frac{B_1^*}{A_1^*}$. It then follows from Eqn. (30) and Eqn. (31) that

$$\begin{aligned} &\frac{\Lambda(\mu_1 r - c - \mu_1 p_1'')}{\mu_1(\beta_1'' p_1'' t_1'' + 1)} - \frac{cM\Lambda}{\mu_1(\mu_1 \beta_1'' p_1'' t_1'' + \mu_1 - \Lambda)} > \\ &\quad \frac{\Lambda(\mu_1 r - c - \mu_1 p_1^*)}{\mu_1(\beta_1^* p_1^* t_1^* + 1)} - \frac{cM\Lambda}{\mu_1(\mu_1 \beta_1^* p_1^* t_1^* + \mu_1 - \Lambda)}, \end{aligned}$$

which is equivalent to

$$\begin{aligned} &\mu_1 \beta_1''^2 (p_1'' t_1'')^2 + (\mu_1 - \Lambda + 1)\beta_1'' p_1'' t_1'' > \\ &\quad \mu_1 \beta_1^{*2} (p_1^* t_1^*)^2 + (\mu_1 - \Lambda + 1)\beta_1^* p_1^* t_1^*. \end{aligned} \quad (32)$$

If we substitute $\beta_1'' = \frac{1}{L_1} \sum_{j \geq 2} At_j'' = \frac{(1+\theta_2)}{L_1} \sum_{j \geq 2} At_j^*$ and $\beta_1^* = \frac{1}{L_1} \sum_{j \geq 2} At_j^*$ into inequality (32), we can obtain the result that

$$\begin{aligned} &\frac{\mu_1 \sum_{j \geq 2} At_j^*}{L_1} [(1 + \theta_2)(p_1'' t_1'' + p_1^* t_1^*) + (\mu_1 - \Lambda + 1)] \\ &[(1 + \theta_2)p_1'' t_1'' - p_1^* t_1^*] > 0. \end{aligned}$$

This implies that $(1 + \theta_2)p_1'' t_1'' > p_1^* t_1^*$, which also implies that $At_1'' < (1 + \theta_2)At_1^*$.

Now let us consider the equilibrium result $\Phi' = (At_1', At_2', ..., At_N')$. Since $\theta_j \leq \theta_2, \forall j > 2$, we have $At_j' \leq At_j'' \forall j \geq 2$. Again, from our proof of Lemma 3, we can obtain the result that $p_1't_1' \geq p_1''t_1''$. Then we will have the following result that

$$At_1'' = L_1 p_1^{''-1} t_1^{''-1} \geq L_1 p_1^{'-1} t_1^{'-1} = At_1' = (1+\theta_1)At_1^*.$$

Combining the two results, it is not difficult to have the following result that

$$(1+\theta_2)At_1^* > At_1'' \geq (1+\theta_1)At_1^*,$$

which means that $\theta_2 > \theta_1$. This contradicts with our assumption that $\theta_1 \geq \theta_2$, and therefore, completes our proof of the uniqueness of the Nash equilibrium. ■

Again, when all cloud providers in the market have the same resource capacity, we are able to obtain an analytical form of the Nash equilibrium solution, which can be derived by solving the optimization problem (24). In the next theorem, we wish to establish the fact that a unique Nash equilibrium exists in the oligopoly market of homogeneous cloud providers, and that Algorithm 2 always converges to such an equilibrium price.

***Theorem** 5:* When reduced to homogeneous cloud providers, *i.e.*, $\mu_i = \mu$ and $L_i = L$, for all i, the Nash equilibrium with N cloud providers $(p_1^*, p_2^*, ..., p_N^*)$ takes the form of:

$$p_i^* = r - \frac{c}{\mu} - \frac{NMc}{\mu(N\mu - \Lambda)} - \frac{NMv}{\Lambda},$$

with the corresponding market shares $f_i = \Lambda/N$, for all i.

Proof: Again, when cloud providers are equivalent to one another, the equilibrium solution is symmetric [26], *i.e.*, $f_i^* = f^*$ and $p_i^* = p^*$ for all i. From the MCI model of the market share, we know immediately that the market share of each cloud provider i equals

$$f_i^* = \Lambda \left(\frac{L_i p_i^{*-a} t_i^{*-b}}{\sum_{j=1}^{N} L_j p_j^{*-a} t_i^{*-b}} \right) = \frac{\Lambda}{N} = f^*,$$

since $t_i^* = \frac{1}{\mu - f^*} = t^*$ for all cloud provider i. Since f_i^* also takes the form of $f_i^* = \frac{\Lambda}{\beta_i^* p_i^* t_i^* + 1}$, we have $\beta_i^* p_i^* t_i^* + 1 = N$. Based on the constraints in problem (24), we have

$$\frac{\Lambda(\mu r - c - \mu p_i)}{\mu N} - \frac{cM\Lambda}{\mu(\mu N - \Lambda)} \geq Mv,$$

which is equivalent to

$$p_i \leq r - \frac{c}{\mu} - \frac{MvN}{\Lambda} - \frac{cMN}{\mu(\mu N - \Lambda)},$$

Now we can complete the proof that with the objective of maximizing the expected profit, the optimal price for each cloud provider i takes the form of $p_i^* = r - \frac{c}{\mu} - \frac{NMc}{\mu(N\mu - \Lambda)} - \frac{NMv}{\Lambda}$, for all i. ■

By comparing with results in Theorem 2, we can see that the Nash equilibrium in an oligopoly market is in the general form of that in a duopoly market. All cloud providers in the market will charge the same price that has the form of the monopoly price p_Λ, with each of them taking $1/N$ of the market. In other words, each cloud provider will behave independently and operate exactly the same as a monopolist, when all of them have the same resource capacity. The comparative statics of the homogeneous Nash equilibrium price in an oligopoly market is the same as what was shown in Corollary 2.

V. Evaluation

We now present our evaluation results based on simulations, on how the Nash equilibrium is influenced by both cloud providers and cloud users. From the cloud providers' perspective, we study the effects of resource capacities on equilibrium prices. On the cloud users' side, we show how the task importance and urgency can influence the equilibrium prices of the cloud service. The design of our simulator is based on a time-slotted synchronous model, with all events generated and processed in their respective time slots. Our simulator is developed in the MATLAB environment.

A. Analyzing the Nash Equilibrium in a Duopoly Market

We begin our evaluation with two cloud providers competing in the market. Since the proposed algorithm is shown to be able to find the Nash equilibrium prices within a small number of iterations, the equilibrium prices in each simulated scenario are obtained by Algorithm 1. Our simulation results have further validated our analytical results in Sec. IV.

We assume that there are 20 cloud users in the market, *i.e.*, $M = 20$. Except otherwise specified, the resource capacity of each cloud provider is set to be $\mu_1 = 2$ and $\mu_2 = 4$; the operating costs $\gamma_1(\mu_1)$ and $\gamma_2(\mu_2)$ are set to be 0 to focus only on the price competition; the reservation value v is set to be 1; the benefit factor r is set to be 5; and the waiting cost factor c is set to be 1 for all cloud users. User j's request rate λ_j for the cloud service is uniformly random in $(0, 0.05)$, as the total request rate has to be smaller than the total service rate to avoid an unlimited queueing delay.

Effects of resource capacities on equilibrium prices: We first study how a cloud provider's resource capacity, μ, will affect the Nash equilibrium prices. In this scenario, μ_1 is set to be 2, $\mu_2 > \mu_1$ and their ratio is assumed to range from 1 to 4. Fig. 6 shows how the Nash equilibrium price and the market share of cloud provider 2 react when its server capacity increases, while that of cloud provider 1 remains the same. As we can see, both the usage price and the market share increase with the resource capacity. To further understand the impact of resource capacities on both cloud providers, we compute the ratio of Nash equilibrium prices as well as the ratio of market shares of the two cloud providers. Our results in Fig. 7 show that, when resource capacities change while other characteristics remain constant, the comparative advantage of cloud provider 2 to cloud provider 1 on both price and the market share also increases, which further proves the importance of the resource capacity.

Effects of resource capacities on expected profits: Since the objective of cloud providers is to maximize their expected profits, we now discuss how expected profits are influenced by resource capacities. In our discussion, the expected profit of a cloud provider equals its usage price times its market share. Shown in Fig. 8, the expected profit of cloud provider 2 does not increase evidently after a certain threshold, and may even decrease, with linear, exponential and step functions for operating costs. The rationale is that when the resource capacity increases, the cloud provider is able to charge a higher price and attract more users, and hence is able to

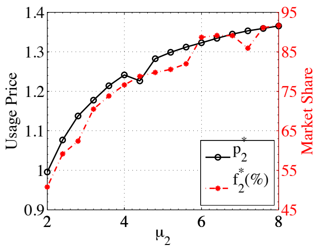


Fig. 6. Effects of resource capacities on a cloud user's equilibrium price and its market share.

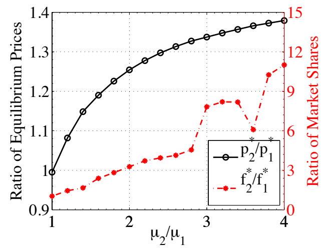


Fig. 7. Effects of resource capacities on the ratio of equilibrium prices and market shares.

generate more profit. However, if the cloud provider continues to increase its resource capacity, the operating cost, such as power consumption, may be too large to be made up by the increased revenue, which may result in reduced profit.

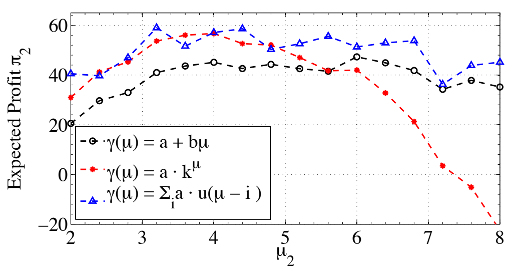


Fig. 8. Effects of resource capacities on a cloud provider's expected profit with different operating cost functions ($a = 10$, $b = 8$, and $k = 1.4$).

B. Analyzing the Nash Equilibrium in an Oligopoly Market

We are now ready to use our algorithm in finding the Nash equilibrium prices to derive some interesting observations, as multiple cloud providers are competing for the same pool of cloud users. Our evaluation is conducted in a scenario with 4 cloud providers, with resource capacities of $\mu_1 = 150$, $\mu_2 = 250$, $\mu_3 = 200$, and $\mu_4 = 500$, respectively. To focus on the price competition only, the operating costs are set to be zero, and the combined effects of other factors are set to be $L_1 = L_2 = L_3 = L_4 = 1$. With the same number of 20 cloud users in the market, the reservation value v is set to be 1; the benefit factor r is set to be 20; and the waiting cost factor c is set to be 100 for all cloud users. User j's request rate λ_j for the cloud service is uniformly random in $(1, 10)$.

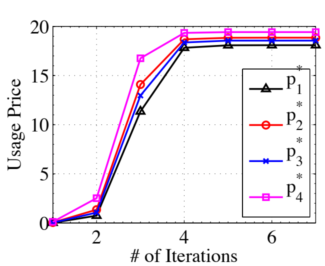


Fig. 9. Finding the equilibrium prices in a cloud market.

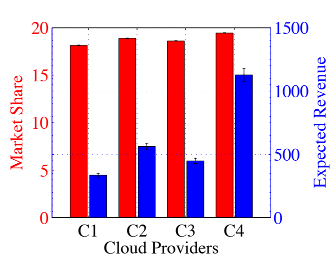


Fig. 10. The Nash equilibrium market shares and expected revenues of each cloud provider.

Fig. 9 shows the price of each cloud provider at each iteration. As we can see, when we set the initial price and the convergence criteria of four cloud providers to be 0.001, our proposed algorithm can find the Nash equilibrium prices in this market within 7 iterations. The Nash equilibrium has further confirmed our conclusion that the cloud provider with a larger resource capacity is able to charge a higher price in a competition market. In addition, Fig. 10 shows the average market shares and expected revenues of each cloud provider, together with the corresponding 95% confidence interval. It is interesting to observe that, though equipped with very different resource capacities (*e.g.,* $C4$ offers twice the capacity over $C2$), each of the four cloud providers takes approximately an equal amount of the market share. This shows that doubling the resource capacity may not necessarily increase the market share by any significant margin. With respect to the expected revenue, a larger resource capacity does lead to more revenue; but this may not directly translate to higher profits, due to the escalating capital and operation expenses associated with larger capacities. We will discuss more implications of these observations in the next section.

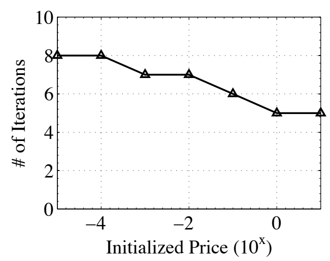


Fig. 11. Effects of the initial price on equilibrium prices.

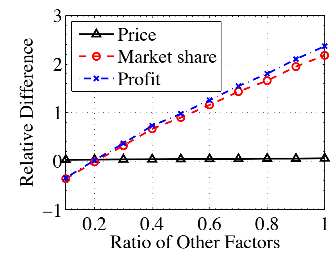


Fig. 12. Effects of the combined attraction of other factors.

Since the initial price of each cloud provider may affect the number of iterations that our algorithm needs to achieve the Nash equilibrium, we are interested in how this initial price will affect the convergence of the proposed algorithm. Shown in Fig. 11, when the initial price varies from 10 to 10^{-5}, the number of iterations our algorithm requires to converge has been very mildly affected, only increasing from 5 to 8. As a result, even if we set the initial price of each cloud provider to be a very small number to avoid missing the Nash equilibrium price, our algorithm can always converge within a few iterations.

As introduced in Sec. IV-B, the Multiplicative Competitive Interaction (MCI) model is used in our analysis to capture the relationship between usage prices and market shares in an oligopoly market. It has incorporated the ability to represent how cloud users feel towards the service of a cloud provider, including alternative influential factors such as whether the Application Programming Interface (API) is secure and easy to use, how load balancing is to be performed, as well as the reputation and brand of the cloud provider. As the final experiment in this section, we are interested in investigating the combined effects of these alternative factors. We use two cloud providers as an example, with resource capacities $\mu_1 = 150$ and $\mu_2 = 500$. Fig. 12 shows when the ratio of $\frac{L_2}{L_1}$ varies from 0.1 to 1, how relative differences of their usage prices, their market shares, and their expected profits change accordingly. As the ratio becomes smaller, it represents the fact that the provider with a larger capacity has become *less attractive* to cloud users due to the combined effects of the alternative factors. As we can observe from the figure, being less attractive to cloud users does not affect the usage prices, as the two cloud providers have the same prices over the entire range of ratios (*i.e.*, the relative difference remains zero). That said, as the provider with a larger capacity has become less attractive due to alternative factors, both its market share and its expected profit decrease significantly, to the point that they can be smaller than its competitor with a third of the capacity. This implies that if a cloud provider wishes to keep its competitive edge, it needs to become more attractive with respect to alternative quality factors, such as its reputation and brand. We will discuss more implications in our concluding remarks.

VI. Discussions and Concluding Remarks

In this paper, we have studied the problem of price competition in a market with multiple IaaS cloud providers. In particular, we have focused on answering the question: when multiple IaaS providers face a common pool of potential users, how should each one of them choose the optimal price that maximizes its own profit? Intuitively, if prices are set to be too high, cloud users will choose alternative cloud providers; but if they are too low, the overwhelming demand for resources from a large number of cloud users may increase the task finishing times, therefore negatively affecting the performance of cloud applications and the utility of cloud users. By modeling each provider as an M/M/1 queue, we analyze this problem using a game theoretic technique in monopoly, duopoly and oligopoly markets. We have derived the sufficient condition for the existence of a Nash equilibrium and propose two iterative algorithms for each provider to find its equilibrium price in the duopoly and oligopoly market, respectively. Our algorithms represent a first step towards designing practical mechanisms to price resources in operational IaaS cloud providers, and are shown to converge quickly.

One important question that is closely related to our analyses and evaluation remains: What an IaaS cloud provider, either an established one or a new player making its market debut, should do to attract new customers and to stay competitive? By analyzing the Nash equilibrium in an oligopoly market where multiple IaaS providers compete, our evaluation have pointed to some intriguing observations that are worth discussing.

At a first glance on our evaluation results, to become more competitive in the market and to gain a larger market share, a cloud provider may initially choose to increase its resource capacity. Yet, the Total Cost of Ownership (TCO), including capital expenses (CAPEX) and operating expenses (OPEX), escalates as cloud providers add to their resource capacities. Such escalating costs may become an important contributing factor that leads to much smaller marginal gains, or even marginal losses, in profits at the IaaS cloud providers.

Our results in Sec. V-B have provided some interesting insights for further discussions on improving the profit of an IaaS cloud provider. As shown in Fig. 10, a larger IaaS provider with twice the resource capacity may only gain a very marginal increase in its market share. Coupled with the higher TCO in maintaining higher capacities, a cloud provider should definitely consider to improve other alternative dimensions of quality in its cloud services, including security, privacy policies, availability of services, data consistency, and ultimately, better service-level agreements that guarantee the performance in a combination of service-level objectives. The aggregate of additional dimensions in its quality of service will be reflected over time in the reputation or brand value of a cloud provider, which will lead to a higher market share and profit.

In Sec. V-B and Fig. 12 in particular, thanks to the fact that our MCI model is able to capture the effects of alternative factors, we have shown that the market share and profit of a much larger cloud provider with more than triple the resource capacity may be similar to its much smaller competitor, if the combined effect of the other alternative quality dimensions is five times worse. This observation provides a strong motivation for an IaaS cloud provider to offer service-level agreements with regards to a number of service-level objectives (*e.g.,* availability and response times) that are weighted more heavily in the cloud user's favour, and to charge higher prices accordingly. While queueing theoretic models are not suitable to provide explicit guidelines on how stronger guarantees may be provided by the IaaS provider on security, privacy, availability, and data consistency, they are nevertheless important factors that contribute to the brand and reputation of the provider.

References

[1] *Amazon EC2 Pricing*. http://aws.amazon.com/ec2/pricing/.

[2] *GoGrid VM Pricing*. http://www.gogrid.com/products/infrastructure-cloud-servers.

[3] *RackSpace VM Pricing*. http://www.rackspace.com/cloud/servers/pricing/.

[4] A. Li, X. Yang, S. Kandula, and M. Zhang, "CloudCmp: Comparing Public Cloud Providers," in *Proc. the 10th Annual Conf. on Internet Measurement (IMC)*, 2010, pp. 1–14.

[5] Y.-J. Hong, J. Xue, and M. Thottethodi, "Dynamic Server Provisioning to Minimize Cost in an IaaS Cloud," in *Proc. ACM SIGMETRICS Joint Int'l Conf. on Measurement and Modeling of Computer Systems*, 2011, pp. 147–148.

[6] K. Tsakalozos, H. Kllapi, E. Sitaridi, M. Roussopoulos, D. Paparas, and A. Delis, "Flexible Use of Cloud Resources through Profit Maximization and Price Discrimination," in *Proc. IEEE 27th Int'l Conf. on Data Engineering (ICDE),*, 2011, pp. 75–86.

[7] V. Kantere, D. Dash, G. Francois, S. Kyriakopoulou, and A. Ailamaki, "Optimal Service Pricing for a Cloud Cache," *IEEE Trans. on Knowledge and Data Engineering*, vol. 23, no. 9, pp. 1345–1358, 2011.

[8] F. Teng and F. Magoules, "Resource Pricing and Equilibrium Allocation Policy in Cloud Computing," in *Proc. IEEE 10th Int'l Conf. on Computer and Information Technology (CIT)*, 2010, pp. 195–202.

[9] M. Mihailescu and Y. M. Teo, "On Economic and Computational-Efficient Resource Pricing in Large Distributed Systems," in *Proc. 10th Int'l Conf. on Cluster, Cloud and Grid Computing (CCGrid)*, 2010, pp. 838–843.

[10] Y. Feng, B. Li, and B. Li, "Bargaining Towards Maximized Resource Utilization in Video Streaming Datacenters," in *Proc. IEEE INFOCOM*, 2012.

[11] H. Xu and B. Li, "Maximizing Revenue with Dynamic Cloud Pricing: The Infinite Horizon Case," in *Proc. IEEE ICC*, 2012.

[12] I. Petri, O. Rana, and G. C. Silaghi, "SLA as a Complementary Currency in Peer-2-Peer Markets," in *Proc. 7th Int'l Conf. on Economics of Grids, Clouds, Systems, and Services (GECON)*, 2010, pp. 141–152.

[13] I. Petri, O. F. Rana, Y. Regzui, and G. C. Silaghi, "Risk Assessment in Service Provider Communities," in *Proc. 7th Int'l Conf. on Economics of Grids, Clouds, Systems, and Services (GECON)*, 2012, pp. 135–147.

[14] H. Chen and M. Frank, “Monopoly Pricing When Customers Queue,” *IIE Transactions*, vol. 36, no. 6, pp. 569–581, 2004.

[15] H. Chen and Y.-W. Wan, “Price Competition of Make-to-Order Firms,” *IIE Transactions*, vol. 35, no. 9, pp. 817–832, 2003.

[16] G. Allon and A. Federgruen, “Competition in Service Industries,” *Operations Research*, vol. 55, pp. 37–55, 2007.

[17] J. Anselmi, U. Ayesta, and A. Wierman, “Competition yields Efficiency in Load Balancing Games,” *Perform. Eval.*, vol. 68, no. 11, pp. 986–1001, 2011.

[18] V. Gupta and R. Nathuji, “Analyzing Performance Asymmetric Multicore Processors for Latency Sensitive Datacenter Applications,” in *Proc. 2010 USENIX Workshop on Power Aware Computing and Systems*, 2010.

[19] J. Bi, Z. Zhu, R. Tian, and Q. Wang, “Dynamic Provisioning Modeling for Virtualized Multi-tier Applications in Cloud Data Center,” in *Proc. 2010 IEEE 3rd Intl. Conf. on Cloud Computing*, 2010.

[20] M. Guevara, B. Lubin, and B. Lee, “Navigating Heterogeneous Processors with Market Mechanisms,” in *Proc. 19th IEEE Intl. Symp. on High-Performance Computer Architecture (HPCA)*, 2013.

[21] O. Gemikonakli, E. Ever, and E. Gemikonakli, “Performance Modelling of Virtualized Servers,” in *Proc. 12th Intl. Conf. on Computer Modelling and Simulation (UKSim)*, 2010, pp. 434–438.

[22] R. W. Wolff, *Stochastic Modelling and the Theory of Queues*. Prentice-Hall, Engelwood, 1989.

[23] L. Rao, X. Liu, L. Xie, and W. Liu, “Minimizing Electricity Cost: Optimization of Distributed Internet Data Centers in a Multi-Electricity-Market Environment,” in *Proc. IEEE INFOCOM*, 2010.

[24] Y. Zhang, Y. Wang, and X. Wang, “GreenWare: Greening Cloud-Scale Data Centers to Maximize the Use of Renewable Energy,” in *Proc. ACM Middleware*, 2011.

[25] M. Mazzucco and D. Dyachuk, “Balancing Electricity Bill and Performance in Server Farms with Setup Costs,” *Future Generation Computer Systems*, 2011.

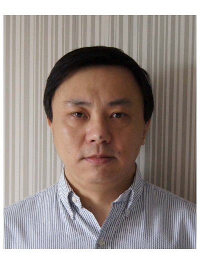
[26] K. C. So, “Price and Time Competition for Service Delivery,” *Manufacturing and Service Operations Management*, vol. 2, pp. 392–409, 2000.

[27] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets — Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[28] D. E. Bell, R. L. Keeney, and J. D. C. Little, “A Market Share Theorem,” *Journal of Marketing Research*, vol. 12, no. 2, pp. 136–141, 1975.

Yuan Feng received her B.Engr. from the School of Telecommunications, Xidian University, Xi’an, China, in 2008, and both her M.A.Sc. and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Toronto, Canada, in 2010, and 2013, respectively. She is currently an Assistant Professor at the Department of Computing, Hong Kong Polytechnic University. Her research interests include optimization and design of large-scale distributed systems and cloud services.

Baochun Li is a Professor at the Department of Electrical and Computer Engineering at the University of Toronto, and holds the Bell University Laboratories Endowed Chair in Computer Engineering. His research interests include large-scale multimedia systems, cloud computing, peer-to-peer networks, applications of network coding, and wireless networks. He is a member of ACM and a senior member of IEEE.



Bo Li is a Professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. He holds a Chang Jiang Visiting Chair Professorship in Shanghai Jiao Tong University (2010-2013), and is the Chief Technical Advisor for ChinaCache Corp., the largest CDN operator in China. His current research interests include content distribution, datacenter networking, cloud computing, mobile applications, and vehicular networks. He received four best paper awards from IEEE, and is a Fellow of the IEEE.