Optimal Pricing for Service Provision in an IaaS Cloud Market with Delay Sensitive Cloud Users

Gang Fang[1] , Xianwei Li[2,3]

[1]**Trade Circulation Institute Anhui Business College Hefei, China**
Email:1511154153@qq.com

[2]**School of Information Engineering Suzhou University Suzhou, China**

[3]**Global Information and Telecommunication Institute Waseda University Tokyo, Japan**
Email:lixianwei163@163.com

Abstract. Cloud computing has received a significant amount of attentions from both engineering and academic fields. Designing optimal pricing schemes of cloud services plays an important role for the success of cloud computing. How to set optimal prices of cloud resources in order to maximize these CSPs' revenues in an Infrastructure as a Service (IaaS) cloud market while at the same time meeting the cloud users' demand satisfaction is a challenging problem that CSPs should consider. However, most of the current works on cloud market are performed under the assumption that cloud users are not sensitive to delay, which is not practical. Towards this end, in this paper we study price-based service provision in an IaaS cloud market. Our simulations verify our analysis

Keywords: Price Competition, IaaS, CSP

1. Introduction

In recent years, cloud computing has received a significant amount of attentions from both engineering and academic fields and the use of cloud service is proliferating. Cloud computing can be defined by several ways, one widely adopted is proposed by Buyya et al. [1] :

> "A cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and the consumers"

Cloud services are mainly classified into three types [2]: Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS). A recent study show that the market size of cloud computing will reach $112 billion in 2018, in a large part due to IaaS cloud services [3]. We focus on IaaS clouds in this paper, where CSPs deliver Infrastructure as a Service (IaaS) to cloud users. In the cloud computing environment, IaaS CSPs bundle their physical resources, such as CPU, memory and disk, into distinct types of virtual machine (VM) instances, according to their sizes and features, and offer them as services to users. Amazon EC2 is a public CSP which has hosted several types of VM instances (e.g. small, medium, large and extra large) based on the capacities of CPU, memory and disk [4], the configurations of some VM instances are shown in Table 1. Cloud users purchase units of computing time on VM instances to run their jobs.

The rapidly increasing demand for cloud resources from business and individuals is making resource management become the heart of CSPs' decision-process [5], and pricing provides an effective approach to addressing this issue. Since the amount of resources that users' request is much smaller than the capacity of CSPs [6], a rational user will subscribe to choose services from the one that maximizes its net reward, i.e., the utility which measures its satisfaction from using cloud service. With more and more IaaS CSPs beginning to provide cloud services, they compete with each other for existing and attract future cloud users. On one hand, CSPs want to charge more from users to maximize their revenues. On the other hand, if they set the prices of cloud services too high, they may have the risk of losing cloud users in the long run. Therefore, how to set the optimal prices to make the revenue maximized while attracting cloud users is a challenging problem, especially when CSPs have different cloud capacities. Furthermore, computing resources, such as CPU cycles and disk, are inherently perishable, that is, they are of no value if they are not utilized in time [7]. In addition, even for the similar type of VM, different CSPs have different prices. For example, although Amazon EC2 m1.medium and Google n1-standard1 have the similar configurations (one virtual CPU and 3.75 GB RAM), they have different prices for one-hour usage.

Recent studies report that different IaaS CSPs process tasks with different completion time [8]. From the perspective of cloud users, besides price quality of service (QoS) is also an important factor that affects the choice of them. Although QoS can be measured by several parameters, such as response time, availability and throughput, all of which can be determined by making use of the tool of queueing theory [9][10]. We mainly focus on response time as the measurement of QoS [8][11].

Table 1 Configurations of Some Amazon EC2 VM Instances

Instance Types	Compute Unit	Storage (GB)	Memory (GiB)
c3.large	2	32SSD	3.75
c3.xlarge	4	80SSD	7.5
c3.2xlarge	8	160SSD	15
c3.4xlarge	16	32SSD	30
c3.8xlarge	32	80SSD	60

A significant amount of works have been devoted to resource management in cloud computing, but only a small fraction of them involved performance issues. In [11], the authors we presented an aggressive virtualized resource management system for IaaS clouds based on reinforcement learning approach. Hong et al. [7] investigated optimal resource allocation for cloud users in an IaaS cloud by developing a dynamic programming algorithm to minimize CSPs' costs. The authors in [3] studied optimal resource allocation in a federated cloud, and they proposed a cloud federation mechanism that enables IaaS CSPs to maximize their profits. Kantere et al. studied the correlation between user demand and the price, and proposed a novel price-demand model to maximize the CSPs' profits [8]. However, these previous works only considered delay-tolerant jobs ignoring delay which is of great important for users who run delay-sensitive jobs. This is because the delayed response time may discourage cloud users to subscribe cloud service or make them switch to other CSPs, which will cause revenue loss. Recent study shows that every 100ms of latency cost Amazon 1 percent in sales and traffic dropped 20 percent if an extra 0.5 seconds happened in search page generation time in Google [11].

Queuing theory are widely adopted to model CSPs' data centres and computing platforms. Feng et al. studied price competition in an oligopoly cloud market with multiple IaaS CSPs, each of which is modelled as an M/M/1 queue [8]. Atmaca et al. proposed a G/G/c-like queuing model to represent a cloud computing system and compute expected performance indices. Their model has the advantage in that it can represent general distributions of workloads on the arrival and service patterns in the cloud computing systems [12]. Khazaei et al. present an approximate model by using an M/G/m/m+K queue with general service time and Poisson arrivals to evaluate the performance of active VM instances [13]. Based on [13], similar model is also adopted by Chang et al. for the study of an IaaS cloud data center [14]. A hierarchical stochastic model is proposed by in [13] to analyze several factors such as variation in job arrival rates and buffer size that affect the quality of cloud service. Most of the aforementioned works are carried out under the assumption that there is an IaaS CSP in the cloud market, which is not realistic as cloud market is becoming more and more fierce with an increasing number of CSPs begin to provision cloud services. Only few works take competition between CSPs into account (such as [8]) restricted to homogeneous cloud markets, that is, theses IaaS CSPs have homogeneous cloud capacities. Without considering users' utilities, the heterogeneous cloud market is originally explored in [15], where the authors analyze the price competition between a public CSP and a cloud broker.

In this paper, we study price competition in a heterogeneous IaaS cloud market by taking CSPs' heterogeneous cloud capacities into consideration. We consider a monopoly cloud market where a resource-constrained CSP modelled as an M/M/1 queuing system offers services to a potential stream of cloud users. Given the price of cloud service, we analyze cloud users' joining policy and show that there exists a unique equilibrium arrival rate to CSP.

2. System Model

In this section, we introduce the models of cloud users' and CSPs. As illustrated in Fig.1, we consider an IaaS cloud computing market with two CSPs to compete for a potential stream of cloud users. CSP1 has constrained cloud resources while CSP2 has sufficient cloud resources, that is, the cloud market is heterogeneous. One example is that CSP1 is an entrant CSP and CSP2 is an incumbent one

2.1 Cloud Users' Model

We assume that the tasks of users arrive at the cloud market with rate Λ following Poisson and they are served according to first-come-first-served (FCFS) queueing. According to recent studies for the analysis of cloud data centers, it is generally accepted that users' service requests arrive at the cloud servers follow Poisson distribution [16]. Similar to [14], we also assume that each job consists of one task, which is single-task job. Each user is supposed to carry a different task, therefore, we use task and user interchangeably. Upon arrival, each cloud user will make a decision to choose from one of the two CSP based on prices and quality of service (QoS) to buy cloud services to execute its task. The jobs of users can be classified into two types [17]: interactive (delay-sensitive) jobs, such as web service, and batch (delay-tolerant) jobs, such as scientific applications. We focus on the study of delay-sensitive jobs. Based on the above assumptions, the utility that a cloud user get from using cloud service of CSP is denoted as

$$U = R - cw(\lambda) - p \tag{1}$$

where R is the reward from using cloud service, w(λ) is the delay time in the cloud system of CSP, c is the delay cost per unit time and p is the per unit time price of VM instance of CSP. Similar utilities functions are widely used in the cloud computing literature [8][15][18].

2.2 CSP model

We model the CSP as an M/M/1 queue whose resource capacity is characterized by service rate μ (in tasks/s) as illustrated in Fig.2.

3. Monopoly Cloud Market

We study a monopoly cloud market, where there is a CSP provisioning cloud services to a potential stream of cloud users. Cloud users arrive at the cloud market with rate Λ. We analyze the relationship between the CSP and users as a two stage Stackelberg game, as illustrated in Figure 3. In the first stage, CSP sets optimal prices to maximize its revenue given the arrival rates of users. In the second stage, cloud users make their arrival rates decision based on the prices of cloud services. The Stackelberg game is solved by using backward induction method [19].

A cloud user' s net utility from using the cloud service is odeled as

$$U = R - p - cw(\lambda) \tag{2}$$

where $w(\lambda) = \frac{1}{\mu - \lambda}$ is the response time includes waiting time and processing time. We assume $\mu > \lambda$ in order to stabilize the queue.

To maximize his utility, a cloud user will pay to use this CSP' s service if

$$U = R - p - cw(\lambda) \geq 0 \tag{3}$$

and refuse to use it otherwise.

Similar to the existing works [8] [20] we consider the equilibrium case, which means

$$R - p - c\frac{1}{\mu - \lambda} = 0 \tag{4}$$

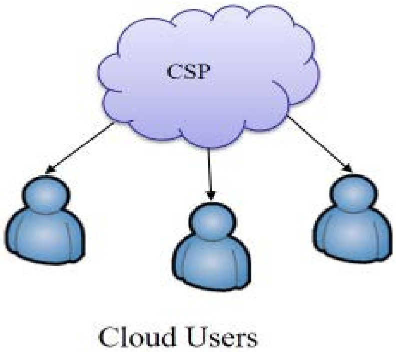

Figure.1 An IaaS cloud marke

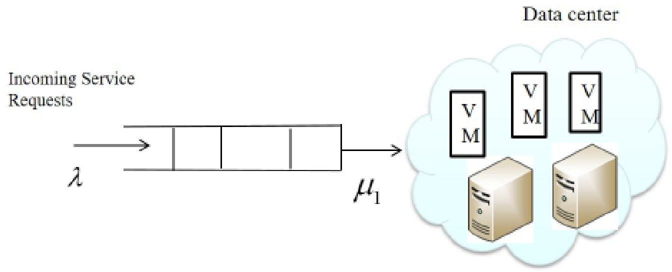

Figure.2 CSP1 is odeled as an M/M/1 queue

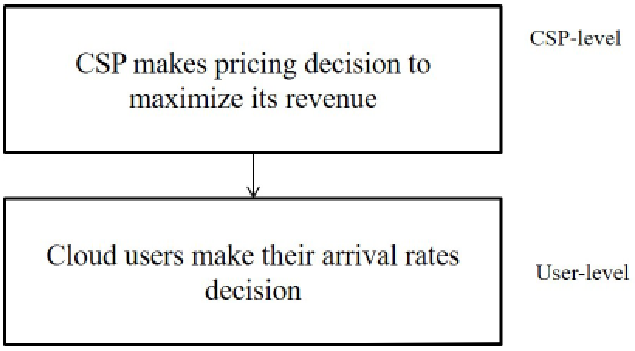
CSP makes pricing decision to maximize its revenue
CSP-level
Cloud users make their arrival rates decision
User-level

Figure.3 A Two-Stage Stackelberg Game

From the Eq. (4), we get

$$\lambda = \mu - \frac{c}{R-p} \tag{5}$$

If the CSP cannot take the whole cloud market in equilibrium, otherwise. So the actual market share of the CSP is

$$\lambda = \min\{\Lambda, \mu - \frac{c}{R-p}\} \tag{6}$$

The revenue of the CSP per unit time is

$$\max_{0<p<R-c\frac{1}{\mu-\lambda}} \pi = p\lambda \tag{7}$$

where λ is given by (6).

The equilibrium price p* is equal to the first-order price, the form of which is

$$p_m^* = R - \sqrt{\frac{cR}{\mu}} \tag{8}$$

with the corresponding market share is

$$\lambda = \min\{\Lambda, \mu - \sqrt{\frac{\mu c}{R}}\} \tag{9}$$

If the CSP can take the entire cloud market, e.g., $\lambda = \Lambda$, then the market capture price is

$$p_{\Lambda} = R - c\frac{1}{\mu - \Lambda} \tag{10}$$

4. Performance Evaluation

In this section, we do simulations to verify our analysis in the previous sections. In particular, we analyze cloud users’ equilibrium arrival rates and CSP’s revenue to several parameters, such as reward values, delay cost and service rate.

4.1 Cloud Users’ Equilibrium Arrival Rates versus Prices

We first analyze how users’ equilibrium arrival rates versus prices of cloud services p. As shown in Figure 4, equilibrium arrival rates not only decrease with increasing values of prices, but also decrease with delay cost value c increasing.

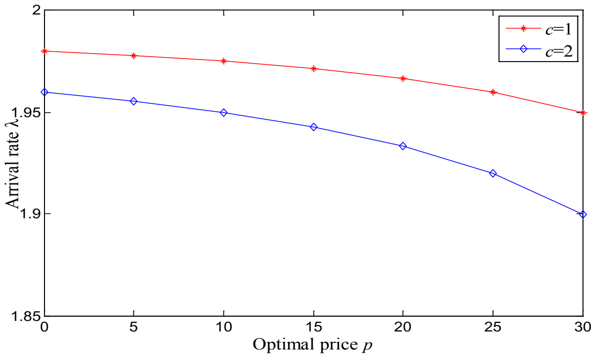

Figure.4 Cloud users’ arrival rate vs optimal price p with r=50, μ =2.

4.2 CSP’s Revenues versus Service Rates

We next analyze how CSP’ s revenues vary with service rates.

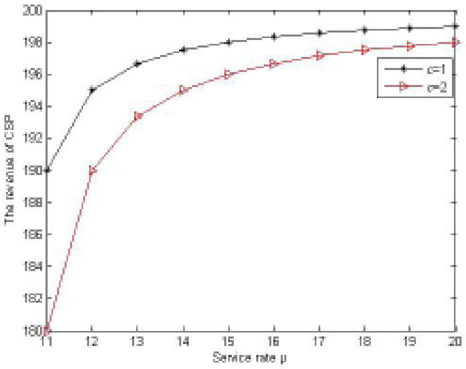

Figure.5 The revenue of CSP versus service rate with R=20, Λ=10.

5. Conclusions

We studied duopoly price competition in an IaaS cloud market in this paper. We model the interactions between CSPs and users a two-stage Stackelberg game, where CSPs set optimal prices to make revenues maximized in the first stage, then cloud users make their arrival rates decision in the second stage. We consider two cloud market cases. The first case is the total arrival rates of the two CSPs is smaller than the market size, and the second case is the total arrival rates of the two CSPs is equal to the market size.

In future works, we will extend our study to duopoly and oligopoly cloud market and study other pricing schemes, such as reservation and spot pricing schemes. We will also study how to segment cloud resources with different pricing schemes.

Acknowledgment

This paper is supported by the following projects, Anhui Key research projects of Humanities and Social Sciences (SK2016A0207), and Suzhou Regional Collaborative Innovation Center (2016szxt05).

References

[1] R. Buyya, C.S. Yeo, and S. Venugopal, “Market Oriented Cloud Computing: Vision, Hype, and Reality for Delivering it Services as Computing Utilities”, Proc. 10th IEEE Conference on High Performance Computing and Communications (HPCC 2008), pp. 5-13, Sept. 2008.

[2] D. Bruneo, “A stochastic model to investigate data center performance and QoS in IaaS cloud computing systems”, IEEE Trans. Parallel Distrib. Syst., vol.25, no.3, pp.560 – 569, March 2014.

[3] L. Zheng, Carlee Joe-Wong, and C. G. Brinton et al. “On the Viability of a Cloud Virtual Service Provider”, Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science (SIGMETRICS 2016), Antibes Juan-les-Pins, France, pp. 235-248, June 2016.

[4] Amazon EC2 Pricing. http://aws.amazon.com/cn/ec2/pricing/.

[5] L. Mashayekhy, M. M. Nejad, and D. Grosu, “Physical machine resource management in clouds: A mechanism design approach”, IEEE Trans. Cloud Comput., vol.3, no.3, pp.247 – 260, July-Sep. 2015.

[6] T.T. Huu and C.K. Tham, “A novel model for competition and cooperation among cloud providers”, IEEE Trans. Comput., vol.2, no.3, pp.251 – 265, July-Sep. 2014.

[7] H.Xu, B.Li, “Maximizing revenue with dynamic pricing: the infinite horizon case”, Proc. IEEE Conference on Communications (ICC 2012), Ottawa, Canada, pp. 2929-2933, June 2012.

[8] Y. Feng, B. Li, and B. Li, “Price competition in an oligopoly market with multiple IaaS cloud providers”, IEEE Trans. Comput., vol. 63, no. 1, pp. 59-73, Jan. 2014.

[9] J. Liu, Y. Zhang, and Y. Zhou et al., “Aggressive resource provisioning for ensuring qos in virtualized environments”, IEEE Trans.Cloud Comput., vol.3, no.2, pp.119 – 131, April-June 2015.

[10] T. Atmaca, T. Begin, and A. Brandwajn et al., “Performance Evaluation of Cloud Computing Centers with General Arrivals and Service ”, IEEE Trans. Parallel Distrib. Syst., vol.27, no.8, pp.2341 – 2348, Aug. 2016.

[11] J. Liu, Y. Zhang, and Y. Zhou et al., “Aggressive resource provisioning for ensuring qos in virtualized environments”, IEEE Trans.Cloud Comput., vol.3, no.2, pp.119 – 131, April-June 2015.

[12] T. Atmaca, T. Begin, and A. Brandwajn et al., “Performance Evaluation of Cloud Computing Centers with General Arrivals and Service”, IEEE Trans. Parallel Distrib. Syst., vol.27, no.8, pp.2341 – 2348, Aug. 2016.

[13] H. Khazaei, J. Misic, and V. B. Misic, “Performance analysis of cloud computing centers using M/G/m/m+r queuing systems”, IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 5, pp. 936 – 943, May 2012.

[14] X. Chang, B. Wang and J. K. Muppala et al., “Modeling Active Virtual Machines on IaaS Clouds Using an M/G/

m/m+k Queue”, IEEE T. Serv. Comput., vol. 9, no. 3, pp. 408 – 420, May 2016.

[15] C.T. Do, N.H. Tran, and E.N. Huh et al., “Dynamics of service selection and provider pricing game in heterogeneous cloud market”, Journal of Network and Computer Applications, vol.69, pp.152 – 165, July 2016.

[16] M. Liu, W. Dou, and S. Yu et al., “A decentralized cloud firewall framework with resources provisioning cost optimization”, IEEE Trans. Parallel Distrib. Syst., vol.26, no.3, pp.621 – 631, March 2015.

[17] Z. Liu, M. Lin, and A. Wierman et al., “Greening geographical load balancing”, IEEE/ACM Trans. Netw., vol.23, no.2, pp.657 – 671, April 2015.

[18] J. Chen, C. Wang, and B. Zhou et al., “Tradeoffs between profit and customer satisfaction for service provisioning in the cloud”, Proc. Of the 20th international symposium on High performance distributed computing (HPDC 2011), San Jose, California, USA, pp.229 – 238, June 2011.

[19] D. Fudenberg and J. Tirole, “Game theory”, MIT Press, Cambridge, MA, USA, 1991.

[20] C. Liu, K. Li, and C. Xu et al., “Strategy Configurations of Multiple Users competition for cloud service reservation”, IEEE Trans. Parallel Distrib. Syst., in press.