Optimal Pricing for Service Provision in IaaS Cloud Markets

Gang Fang
Trade Circulation Institute, Anhui Institute of International Business, Hefei, China, 230000;

Zhengce Cai
Department of Information Service,
Anhui Institute of International Business,
Hefei, China, 230000

Xianwei Li*
School of Information Engineering,
Suzhou University, Suhou, China, 234000;
*lixianwei163@163.com

Abstract —Pricing plays an important role for service provision in cloud computing. In this paper, we investigate price based resource access control in two Monopoly IaaS cloud market, respectively. The two IaaS cloud market is formed by one public cloud service providers (CSPs) and cloud broker (CB), provisioning cloud services to delay-sensitive cloud users (CUs). In the first monopoly cloud market, we treat the public CSP as an M/M/1 queueing system and study this CSP's pricing effect on the equilibrium behaviours of self-interested CUs. We propose two pricing mechanisms with the objective of maximizing revenue and social welfare, respectively. In the second monopoly cloud market, the CB is modelled as an M/M/∞ queueing system, which has infinite capacity to serve a common pool of CUs. We also analyze how pricing affects the equilibrium behaviors of CUs and the revenue-optimal and social-optimal pricing strategies in view of this CSP.

Keywords-Pricing, IaaS; cloud market; queueing system

I. INTRODUCTION

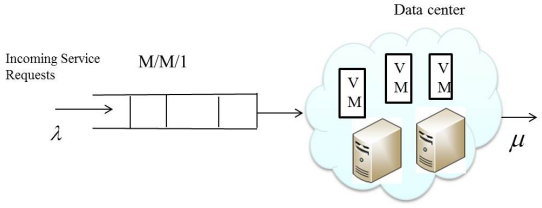
In recent years, cloud computing has received a significant amount of attentions from both engineering and academic fields and the use of cloud service is proliferating. Cloud computing can be defined by several ways, one widely adopted is proposed by Buyya et al. [1] :

"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.

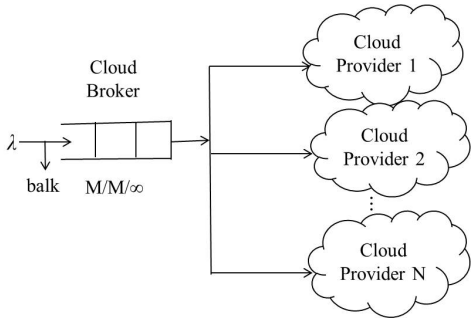
Optimal pricing for cloud resources has been extensively studied by a significant amount of works in the literature. Feng et al. studied non-cooperative price competition in an oligopoly public cloud market [5]. They modelled each PCP as an M/M/1 queue, and analyzed how to set optimal prices in order to maximize the revenues of PCs based on resource capacities and the job finishing time. Xu et al., presented a study pricing cloud resources in a monopoly public cloud market [6]. Their study indicated that the revenue got with reserved pricing is not less than the first-order discrimination pricing. Mashayekhy et al. proposed a federation formulation game that considers the cooperation of these cloud providers to offer cloud services [7]. Their designed cloud federation mechanism enables cloud providers dynamically to form a federated cloud, which maximizes the profits of cloud providers.

In this paper, we study pricing-based service access control of CUs in a heterogeneous cloud market formed by two CSPs, public CSP and CB provisioning cloud services to delay-sensitive CUs. We consider two cloud scenarios corresponding to two types of cloud market: public CSP monopoly, and CB monopoly, which is illustrated in Figure 1. We note that similar structure analysis is also adopted by [9] in which the authors studied optimal pricing effects on the equilibrium behaviours of secondary users in cognitive radio networks. However, the effects of delay costs charged by CSPs on cloud users are not fully considered in [9]. By incorporating the delay costs of CSPs, in the first monopoly



(a) Public cloud monopoly market

978-1-5386-3981-8/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCNEA.2017.18



(b) Cloud broker monopoly market

Figure 1. Two cloud market scenarios

Cloud market, we model the PC as an M/M/1 queueing system and analyze the pricing effect of this CSP on the equilibrium behaviours of non-cooperative delay-sensitive CUs. These behaviours are characterized by CUs' service access decisions of joining or balking to the queue upon arrival. From the viewpoint of CUs, their service access decision model are made according to the individual optimal strategy exploited by each CU, which is based on a utility function that captures the heterogeneous delay- sensitivity of CUs. We then show that there is a unique Nash equilibrium of CUs' joining probability in the non-cooperative game among them. In terms of the monopoly CSP, we design two pricing policies with the objective of maximizing revenue and social welfare, respectively.

In the second monopoly market, the CB is modelled as an M/M/∞ queueing system provisioning cloud services to delay-sensitive CUs. Similar to the first monopoly cloud market, we also study the CSP's pricing effect on the equilibrium behaviours of CUs. Since the CB has sufficient resources to serve the needs of CUs, therefore, it can provide better quality of service (QoS) measured by the average queueing delay. From the perspective of this CSP, we also study two pricing policies with the objective of maximizing revenue and social welfare, respectively.

The rest of the paper is structured as follows. System models are presented in section 2. We analyze the monopoly public cloud market in the section 3, the monopoly CB cloud market in the section 4. Conclusions and future works are given in section 5.

TABLE I. 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

II. SYSTEM MODELS

A. CUs model

We assume that there is potential stream of CUs arrive at the cloud market with rate λ according to the Poisson process. Each CU carries a distinct job upon arrival. Therefore, we use CU and job interchangeably throughout the paper. The jobs of CUs in cloud data centers are classified into two types [10]: interactive (delay-sensitive) jobs, such as web service, and batch (delay-tolerant) jobs, such as scientific applications. Recent study shows that delay-sensitive interactive workloads take over 50% of data center workloads [11]. Hence, we focus on delay-sensitive interactive jobs and assume that each job attached to a specific application is denoted by a parameter θ, which reflects the sensitivity of CU's application to delay. The value of θ is private, but its distributions are known to CSPs. We also assume that θ is uniformly distributed on [0,1] with probability distribution function (PDF) f(.) and cumulative distribution function (CDF) F(.). This assumption is also widely adopted in the literature [9] [12][13].

When a type-θ CU arrives to the cloud market, it must make a decision as to whether to acquire service or not. If joins CSPi (i=p or c, where p and c denotes public CSP and CB, respectively), it will get net utility which is

$$U_i = r - \theta d_i - p_i, i = p, c \tag{1}$$

This net utility function is commonly used in the cloud and communication networks literatures [5][9][13], which captures the balance between the reward r and the total costs θcdi+pi that a CU takes if it joins the queueing system CSPi. The reward r represents the benefit factor of a CU for accessing the cloud service [9][13]. The total costs include two parts: θdi and pi, where di is the average queueing delay that this job experiences in the queueing system and pi is the price per service request charged by this CSPi. The similar pricing scheme is widely adopted by CSPs. Such as, Campaign Monitor [14] and Amazon Simple Email Service (ES) [15] charge CUs according to the number of campaigns and emails they process, respectively.

B. Public Cloud Service Provider (CSP)

When a type-θ CU decides to subscribe the service from the public provider, it will join a queueing system of this public provider. The system of the PC is modelled as an M/M/1 queue with service rate μ serving a potential number of CUs. The M/M/1 queue model is widely used in the cloud computing literature [9] to analyze response time as a function of the capacity of cloud resources and arrival rate of service requests. From (1), the net utility of type-θ CU for accessing the service public provider given price p1 is

$$U_1 = r - \theta c d_1 - p_1 \tag{2}$$

Where $d_1(\lambda) = 1/(\mu - \lambda)$ is the average queueing delay incurred by the arrival rate λ.

C. Cloud Broker (CB)

Since the CB integrates and coordinates resources among different CSPs, therefore, we assume that it has sufficient cloud resources to meet the demands of CUs. Hence, the system of CB is modelled as an M/M/∞ queue with enough servers to serve a common potential pool of CUs. The similar models have been widely used in the cloud literature to analyze power management or resource allocation in data centers. In [16], the authors studied optimal multi-server configuration to maximize profit of CSPs in cloud data centers. In [17], by modelling the CSP as M/G/m/m+r queueing system, the authors analyzed the performances of cloud data centers. Fang et al. studied throughput and energy tradeoff in mobile cloud platforms by applying the M/M/m queueing model [18]. From (1), the net utility of type-θ CU for accessing the service cloud broker given price p2 is

$$U_2 = r - \theta d_2 - p_2 \tag{3}$$

where $d_2 = 1/\mu$ captures the average queueing delay in M/M/∞ queue.

III. PUBLIC CLOUD MONOPOLY MARKET

In this section, we first investigate the decisions of CUs as to whether to join or balk to the public provider and then design two optimal pricing mechanisms with the aim of maximizing revenue and social welfare, respectively.

A. CUs' Decision Policy

We consider a number of CUs arriving at the public cloud market, and these CUs are rational decision-makers in that they are only concerned with their own net utilities. Upon arrival, each type-θ CU has to make a decision whether to join or balk the queueing system of the public provider. It will join the queue if and only if its net utility U1(θ)≥0. Therefore, we get the following individual optimal decision policy.

Definition 1. A self-optimizing type-θ CU with its net utility $U_1(\theta) = r - \theta c d_1(\lambda_1) - p_1$ will follow a joining decision policy such that

- it joins public provider if U1(θ)≥0，which requires θ≤θ1, where

$$\theta_1 = \frac{r - p_1}{c d_1(\lambda_1)} \tag{4}$$

- it balks, if U1(θ)<0.

The above definition indicates that the fraction of CUs that have θ values less than θ1 will subscribe to the public provider. The fraction of CUs that have θ values less than θ1 is

$$F(\theta_1) = \int_0^{\infty} f(\theta) d\theta = \int_0^{\theta_1} f(\theta) d\theta \tag{5}$$

Then, the effective arrival rate of CUs to the public provider denoted by λ1 is

$$\lfloor 1 = \lfloor \Phi(\backslash 1) \tag{6}$$

B. Revenue Optimal Pricing Mechanism

Under the assumption that the public cloud provider knows the effective arrival rate, when charging p1 and delay cost c, this public cloud provider can get revenue $\pi_1(p_1) = p_1\lambda_1$. The objective of the public cloud provider is to maximize its revenue, which can be formulated as

$$\mu\alpha\xi\ \pi_1(p_1) = p_1\lambda_1 \tag{7}$$

$$\text{s.t. } p_1 \in [p_{\text{low}}, p_{\text{up}}]$$

where pup=r, plow=max{0,r-cd(λ1)}.

It is obvious that π1= p1λ1 is a concave function from $\pi_1''(p_1) < 0$. Hence, the problem of (7) can be solved by efficiently. By setting the first derivative $\frac{\partial \pi}{\partial p_1} = 0$, we get the optimal price

$$p_1^* = \frac{c + \lambda r - \sqrt{c(c + r\lambda)}}{\lambda} \tag{8}$$

Accordingly, the optimal revenue is

$$\pi_1^* = \lambda_1 p_1^* = \frac{\mu[2c + \lambda r - 2\sqrt{c(c + r\lambda)}]}{\lambda} \tag{9}$$

C. Social Welfare Optimal Pricing Mechanism

Cloud social welfare is the net utilities of CUs plus the revenue of the public cloud provider. When charging price p1 , only the fraction of CUs with θ≤θ1 subscribe to the public loud provider. Therefore, the cloud social welfare at price p1 is

$$\begin{aligned} S_1(p_1) &= U_1 + \pi_1 \\ &= \int_0^{\theta_1} [r - \theta c d_1(\lambda_1)] f(\theta) d\theta \\ &= r\theta_1 - \frac{\theta_1^2 c}{2(\mu - \lambda\theta_1)} \end{aligned} \tag{10}$$

where θ1 is given in (4). From (4) we know that θ is the function of price p1. Therefore, the variable of $S_1(\theta_1)$ can be changed from p1 to critical CU variable. Hence, the social welfare optimal pricing problem is formulated as

$$\mu\alpha\xi\ S_1(\theta_1) \tag{11}$$

s.t. $\theta_1 \in [0, 1]$

where $S_1(\theta_1)$ is given in (10).

We find that the objective function of problem (11) is concave by calculating $S_1''(\theta_1) < 0$, therefore, the optimal solution of (11) can be effectively solved, which is denoted by θ_1^S. Hence, the optimal social welfare price is

$$p_1^S = r - \theta_1^S c d_1(\lambda \theta_1^S) \tag{12}$$

IV. CLOUD BROKER MONOPOLY MARKET

In this section, we first investigate the decisions of CUs as to whether to join or balk to the cloud broker and then design two optimal pricing mechanisms with the goal of maximizing revenue and social welfare, respectively.

A. CUs' Decision Policy and Equilibrium

We consider a number of CUs arriving at the federated cloud market, and these CUs are rational decision-makers in

f CUs that have θ values less than θ2 will subscribe to the cloud broker. The fraction of CUs that have θ values less than θ2 is expressed as

$$F(\theta_2) = \int_0^{\infty} f(\theta) d\theta = \int_0^{\theta_2} f(\theta) d\theta \tag{14}$$

Then, the effective arrival rate of CUs to the public provider denoted by λ2 is

$$\lfloor 2 = \lfloor \Phi(\backslash 2) \tag{15}$$

B. Revenue Optimal Pricing Mechanism

Under the assumption that the cloud broker knows the actual arrival rate of CUs, when charging p2 and delay cost c, this CSP can get revenue $\pi_2(p_2) = p_2 \lambda_2$. The objective of the cloud broker is to maximize its revenue, which can be formulated as

$$\mu\alpha\xi \ \pi_2(p_2) = p_2 \lambda_2 \tag{16}$$

s.t. $p_2 \in [0, r]$

By setting the first derivative of the objective function with respect to p2 to zero, we get the revenue optimal price

$$p_2^* = \frac{r}{2} \tag{17}$$

Accordingly, the optimal revenue is

ation Center (2016szxt05).

that they are only concerned with their own net utilities. Upon arrival, each type-θ CU has to make a decision whether to join or balk the queueing system of the cloud broker. For a CU, it will join the queue if and only if its net utility U2(θ)≥0. Therefore, we get the following individual optimal decision policy.

Definition 2. A self-optimizing type-θ CU with its net utility $U_2(\theta) = r - \theta c d_2(\lambda_2) - p_2$ will follow a joining decision policy such that

• it joins public provider if U2(θ)≥0，which requires θ≤θ2, where

$$\theta_2 = \frac{r - p_2}{c d_2} \tag{13}$$

• it balks, if U2(θ)<0.

The above definition indicates that the fraction o

$$\pi_2^* = \frac{\mu r^2 \lambda}{4c} \tag{18}$$

C. Social Optimal Pricing Mechanism

The cloud social welfare is defined as

$$\begin{aligned} S_2(p_2) &= U_2 + \pi_2 \\ &= \int_0^{\theta_2} [r - \theta c d_2] f(\theta) d\theta \\ &= r\theta_2(p_2) - \frac{c(\theta_2(p_2))^2}{2\mu} \end{aligned} \tag{19}$$

The cloud social welfare problem is formulated as

$$\mu\alpha\xi \ S_2(p_2) \tag{20}$$

s.t. $p_2 \in [0, r]$

By setting the first derivative of the objective function with respective to p2, the socially optimal price is

$$p_2^S = r - c\mu \tag{21}$$

ACKNOWLEDGMENT

This paper is supported by the following projects, Anhui Key research projects of Humanities and Social Sciences (SK2016A0207), and Suzhou Regional Collaborative Innov

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] 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.

[6] H. Xu and B. Li, “A study of pricing for cloud resources,” ACM SIGMETRICS Performance Evaluation Review, vol. 40, no. 4, pp. 3-12, Mar. 2013.

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

[8] Z. Liu, Y. Chen, and C. Bash, et al., “Renewable and cooling aware workload management for sustainable data centers,” Proc. of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems (SIGMETRICS 2012), London, England, UK, pp. 11–15, June 2012.

[9] N. H. Tran, C. S. Hong, and S. Lee et al., “Optimal Pricing Effect on Equilibrium Behaviors of Delay-Sensitive Users in Cognitive Radio Networks,” IEEE J. Sel. Areas Commun., vol. 31, no. 11, pp. 2266–2579, Oct. 2013.

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

[11] Y. Jin, S. Sen, and R. Guerin, et al., “Dynamics of competition between incumbent and emerging network technologies,” in Proc. Workshop on the Economics of Networks, Systems, and Computation (NetEcon’ 08), Seattle, WA, USA, pp. 49–54, Aug. 2008.

[12] R. Gibbens, R. Mason, and R. Steinberg, “Internet service classes under competition,” IEEE J. Sel. Areas Commun., vol. 18, no. 12, pp. 2490–2498, Dec. 2000.

[13] Campaign Monitor. http://www. campaignmonitor.com/pricing

[14] Amazon SES. https://aws.amazon.com/ses /pricing.

[15] J. Cao, K. Hwang, and K. Li, et al., “Optimal multiserver configuration for profit maximization in cloud computing,” IEEE Trans. Parallel Distrib. Syst., vol.24, no.6, pp.1087-1096, June 2013.

[16] 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.

[17] W. Fang, Y. Li, and H. Zhang, et al. "On the throughput-energy tradeoff for data transmission between cloud and mobile devices." Information Sciences, 283, pp.79-93, 2014.

[18] Fudenberg and J. Tirole, Game Theory, MIT Press, Cambridge, USA, 1991.