Competitive Bandwidth Reservation via Cloud Brokerage for Video Streaming Applications

Xin Jin and Yu-Kwong Kwok
The University of Hong Kong, Hong Kong SAR
{*tojinxin, ykwok*}*@eee.hku.hk*

1. Introduction

The Infrastructure-as-a-Service (IaaS) view of cloud computing is widely adopted by several large cloud providers, which has fundamentally changed the operation of many industries [1-3]. Indeed, large cloud providers such as Amazon Web Services [4], Windows Azure [5**Error! Reference source not found.**], and Google App Engine [6] offer Internet-scale distributed computing facilities, where tenant users can dynamically reserve cloud resources including CPU, memory, and bandwidth so as to satisfy their own service requirements [7].

In such a multi-tenant cloud computing environment, cloud brokers exploit demand correlation among tenants and obtain volume discounts from cloud providers via tenant demand aggregation. Therefore, tenants dynamically procure resources via cloud brokerage services due to lower offered price rates. Therefore, we consider resource procurements from cloud brokers, and tackle the problem of tenant demand competition with a realistic broker pricing policy. In a practical cloud market, resource demands and prices will be cleared at an equilibrium level, where tenant consumers maximize their surplus and cloud brokers optimize the collected revenue given optimal demand responses of tenant consumers.

In this paper, our specific contributions are three-fold. Firstly, we build a general game model to realistically capture broker pricing scheme design. Tenant surplus (i.e., tenant utility minus dollar cost) is realistically formulated to model tenant rationality. Secondly, to relax the impractical assumption of complete information, we propose a dynamic game based bounded rationality to attain Nash equilibrium in a distributed manner by using only local information. Thirdly, we present evaluation results to validate our analytical model and obtain insightful observations.

2. Game Model for Tenant Competition

We consider a cloud system with multiple cloud brokers and a large number of tenant users. Denote by N the number of tenant users in the cloud system. The number of cloud brokers is M. The broker i sells the cloud resources at price rate p_i.

Pricing Model.

The commodity sold in the cloud market is in the units of bandwidth. To model prices offered by cloud broker i, we consider a realistic pricing function where demands affect prices:

$$p_i(\mathbf{d}_i)=\alpha+\beta\cdot\left(\sum_{j=1}^{N} d_{ij}\right)^{\tau},\ \forall i\in\{1,\cdots,M\}, \qquad (1)$$

where d_{ij} is the amount of resources reserved by tenant j from cloud broker i, and $\mathbf{d}_i=[d_{i1},\cdots,d_{ij},\cdots,d_{iN}]^T$ is the vector of all resource demands at broker i. This practically reflects the situation that the price increases with the growth of aggregate demand at one cloud broker due to the limited amount of cloud resources reserved from cloud providers.

Tenant Surplus.

Denote by l_{ij} the network delay due to tenant j's resource procurements from cloud broker i. L represents the maximum experienced network delay in the entire cloud system. Then, the utility of unit bandwidth resource can be modeled as

$$b_{ij}=\ln\left(1+\left(L-l_{ij}\right)\right), \qquad (2)$$

where $L\geq l_{ij}$ and L represents the maximum tolerated delay by tenant consumers. Then, the total utility obtained by tenant user j is $\sum_{i=1}^{M} b_{ij}\cdot\ d_{ij}$, with the financial cost of $\sum_{i=1}^{M} b_{ij}\cdot\ p_i(\mathbf{d}_i)$. Therefore the surplus of tenant j can be formulated as follows:

$$\begin{aligned}\pi_j(\mathbf{s}_j) &= \sum_{i=1}^{M} b_{ij}\cdot\ d_{ij}-\sum_{i=1}^{M} d_{ij}\cdot\ p_i(\mathbf{d}_i)\\ &=\sum_{i=1}^{M} b_{ij}\cdot\ d_{ij}-\sum_{i=1}^{M} d_{ij}\cdot\left(\alpha+\beta\cdot\left(\sum_{j=1}^{N} d_{ij}\right)^{\tau}\right),\end{aligned}$$

where $\mathbf{s}_j=[d_{1j},\cdots,d_{ij},\cdots,d_{Mj}]^T$ is a vector of tenant user j's demands from all the cloud brokers.

Static Game and Nash Equilibrium.

Based on the tenant surplus formulation in the above, we can formulate a non-cooperative game among competing tenant users. The players in this game are all the tenant users. The strategy of each player (e.g., tenant user j) is the demand vector of resources reserved from different cloud brokers (i.e., $\mathbf{s}_j$ for tenant j). The payoff of each tenant user j is the surplus earned from the usage of cloud resources (i.e., $\pi_j(\mathbf{s}_j)$). We use Nash equilibrium to solve the game. The Nash equilibrium of a game is a solution concept in which no player can increase his own payoff by unilaterally changing its own strategy. The Nash equilibrium can be obtained by solving the best response function, which is the optimal strategy of one player given the others' strategy choices. That is, the best response function of tenant j can be formulated as:

$$BR_j(\mathbf{S}_{-j})=\arg\max_{\mathbf{s}_j}\pi_j(\mathbf{S}), \qquad (4)$$

where $\mathbf{S}=[d_{ij}], \forall 1\le i\le M, \textit{ and } 1\le j\le N$ denotes the strategy matrix of all tenant users and $\mathbf{S}_{-j}=[d_{ik}]$ with $i\ne j$ represents the strategy matrix of all tenants except tenant j. To this end, we can obtain the Nash equilibrium by solving the following equation array:

$$\frac{\partial \pi_j(\mathbf{s}_j)}{\partial d_{ij}} = -\beta\cdot\tau\cdot\sum_{i=1}^{M} d_{ij}\cdot\left(\sum_{j=1}^{N} d_{ij}\right)^{\tau-1} = 0. \qquad (5)$$

In the following theorem, we investigate the analytical solution of Nash equilibrium for the special case of M=1. That is, $b_{ij}=b_j$ and $d_{ij}=d_j$, $\forall i$.

THEOREM 1 For the special case of M=1, there exists a unique Nash equilibrium given by

$$d_j^*=\left(\frac{b_j-\alpha}{\beta\cdot\tau\cdot Q^{\tau-1}}-\frac{Q}{\tau}\right)^+, \forall 1\le j\le M, \qquad (6)$$

where $Q=\frac{\sum_{j=1}^{N} b_j-\alpha\cdot N}{\beta\cdot(N+\tau)}$ and $(x)^+=\max(x,0)$.

Proof. From Equation array 5, we get

$$\frac{\partial \pi_j(\mathbf{s}_j)}{\partial d_j} = b_j-\alpha-\beta\cdot\left(\sum_{j=1}^{N} d_j\right)^{\tau} -\beta\cdot\tau\cdot d_j\cdot\left(\sum_{j=1}^{N} d_j\right)^{\tau-1} = 0. \qquad (7)$$

Summing up the left side and the right side of the above equations, we have

$$\sum_{j=1}^{N} b_j-\alpha\cdot N-\beta\cdot N\cdot\left(\sum_{j=1}^{N} d_j\right)^{\tau}-\beta\cdot\tau\cdot\left(\sum_{j=1}^{N} d_j\right)^{\tau}=0.$$

Suppose that $Q=\sum_{j=1}^{N} d_j$. We can readily get

$$Q=\left(\frac{\sum_{j=1}^{N} b_j-\alpha\cdot N}{\beta\cdot(N+\tau)}\right)^{1/\tau}. \qquad (9)$$

Substitute Q into Equation 7, we obtain the unique Nash equilibrium:

$$d_j=\frac{b_j-\alpha}{\beta\cdot\tau\cdot Q^{\tau-1}}-\frac{Q}{\tau}. \qquad (10)$$

However, this is on that condition that

$$d_j=\frac{b_j-\alpha}{\beta\cdot\tau\cdot Q^{\tau-1}}-\frac{Q}{\tau}\ge 0; \qquad (11)$$

otherwise, the best response of tenant j is d_j=0. To sum it up, we obtain the unique Nash equilibrium:

$$d_j^*=\max\left(\frac{b_j-\alpha}{\beta\cdot\tau\cdot Q^{\tau-1}}-\frac{Q}{\tau},0\right). \qquad (12)$$

Dynamic Game and Stability Analysis.

In a practical cloud system, one tenant user may not be aware of the strategies and surplus of the other tenant users. Therefore, each tenant user has to learn others' strategies and pricing behaviors based on the interaction history. To this end, we propose distributed learning algorithms for dynamic demand adjustments so as to gradually achieve Nash equilibrium for competitive resource procurements. In tenant demand competition, tenant users can adjust the resource demands from different cloud brokers towards the most promising direction (i.e., the direction of marginal profit function). Therefore, the adjustment of the optimal demand level is calculated in a dynamic game for tenant j:

$$d_{ij}(t+1)=d_{ij}(t)+\delta_j\cdot d_{ij}(t)\cdot\frac{\partial \pi_j(\mathbf{s}_j)}{\partial d_{ij}}=\Gamma\left(d_{ij}(t)\right), \ \forall i, \forall j, \qquad (13)$$

where $d_{ij}(t)$ is the demand of tenant j from cloud broker i at time slot t and δ_j is the strategy updating step size (i.e., the learning rate) of tenant j. $\Gamma\left(d_{ij}(t)\right)$ is the self-mapping function of the dynamic game. The dynamic game defined by Equation 13 is proposed under the notion of bounded rationality where the tenant users cannot adapt their strategies to the optimal demand levels immediately.

3. Performance Evaluation

In this section, we present our evaluation results. We consider a cloud system with one cloud broker and two tenant users procuring bandwidth from the broker (i.e., M=1 and N=2). In the pricing model, we use α=0 and β=1. The impact of τ is explored by varying its values. By default, we have τ=1. For tenant surplus, we have the maximum incurred delay L=30000. The impact of network delay is examined by varying l_{ij}.

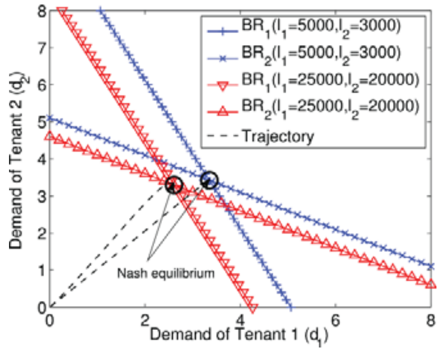


Figure 1: Illustration of Nash equilibrium with two tenant users: best response functions.

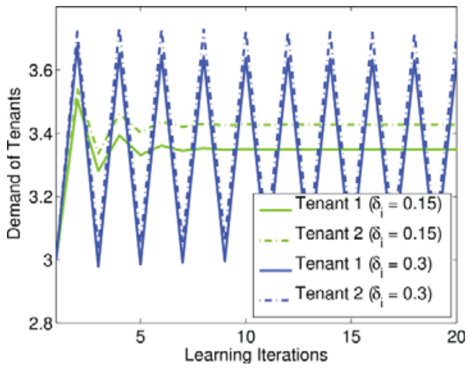


Figure 2: The impact of learning rate on the convergence of the dynamic game.

We first examine Nash equilibrium and the impact of network delay in Fig. 1 for the special case of two tenant users. Here, we investigate the impact of network delay on the equilibrium demand levels. With the decrease of network delay (i.e., better service quality), the corresponding tenant user would like to procure more resources from the cloud broker. On the other hand, the network delay of one tenant user affects the other's procurement of cloud resources. This clearly explains the impact of network delay and the interactions among tenants for resource procurements, when a large number of tenants coexist in the cloud system. We also show the trajectories of the competitive strategies learning of the tenant users in Fig. 1 for the special case of δ_i=0.05 . It shows the convergence of the dynamic game in distributed learning. Fig. 2 shows that, when learning rate is large (e.g., 0.3), the dynamic game may never converge.

4. Related Work

Pricing has been discussed for more than a decade by computer scientists for network resource allocation [8]. Recently, cloud resource pricing is widely adopted as the dominant resource allocation scheme in a cloud computing environment with multi-tenancy. Therefore, there already exist some studies on pricing scheme design and tenant resource procurements. Wang *et al.* [9] examine the importance of cloud resource pricing from the perspective of economics. Due to the coexistence of spot pricing and usage based pricing, Wang *et al.* [10] investigate optimal data center capacity segmentation between both pricing schemes with the objective of total cloud revenue maximization. Niu *et al.* [11, 12] propose a pricing scheme to better leverage the demand correlation among tenant consumers with VoD traffic and argue the necessity of brokers in a free cloud market. Most recently, Xu *et al.* [13, 14] propose centralized schemes so as to maximize the revenue of the cloud provider. Wang *et al.* further discuss optimal resource reservation with multiple purchasing options in IaaS clouds in [15]. While the above studies acknowledge the dominant role of the cloud provider and brokers in pricing, they ignore the competitive cloud resource procurements and its impact on broker revenue and pricing, which is the key problem we aim to solve in this paper.

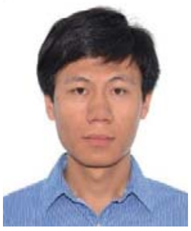
5. Conclusion

In this paper, we explore the problem of competitive cloud resource procurements in a cloud broker market. We realistically model the pricing scheme of the cloud broker and tenant surplus. We propose a non-cooperative game to model such competitive resource procurements. We then conduct equilibrium analysis under the assumption of perfect information. To relax the assumption of perfect information, we propose the adoption of dynamic game to reach Nash equilibrium in a distributed manner by using local information only. The results revealed insightful observations for practical pricing scheme design. In the future, we would like to extend our model to the more general case of an interrelated market formulated by the cloud provider, brokers, and tenant consumers with strategic interactions.

References

[1] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang,

"Multi-Resource Allocation: Fairness-Efficiency Tradeoffs in a Unifying Framework," in Proc. of INFOCOM, March 2012.
[2] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron, "Towards Predictable Datacenter Networks," in Proc. of SIGCOMM, August 2011.
[3] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Greening Geographical Load Balancing," in Proc. of SIGMETRICS, June 2011.
[4] Amazon EC2, 2013, http://aws.amazon.com/ec2/.
[5] Windows Azure Pricing Calculator, 2013, http://www.windowsazure.com/en-us/pricing/calculator/.
[6] Google App Engine, 2013, https://appengine.google.com/.
[7] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, "Dominant Resource Fairness: Fair Allocation of Multiple Resource Types," in Proc. of USENIX NSDI, March 2011.
[8] S. Shenker, D. Clark, D. Estrin, and S. Herzog, "Pricing in Computer Networks: Reshaping the Research Agenda," SIGCOMM Computer Communication Review, vol. 26, no. 2, pp. 19–43, April 1996.
[9] H. Wang, Q. Jing, R. Chen, B. He, Z. Qian, and L. Zhou, "Distributed Systems Meet Economics: Pricing in the Cloud," in Proc. of USENIX HotCloud, June 2010.
[10] W. Wang, B. Li, and B. Liang, "Towards Optimal Capacity Segmentation with Hybrid Cloud Pricing," in Proc. of ICDCS, June 2012.
[11] D. Niu, C. Feng, and B. Li, "A Theory of Cloud Bandwidth Pricing for Video-on-Demand Providers," in Proc. of INFOCOM, March 2012.
[12] D. Niu, C. Feng, and B. Li, "Pricing Cloud Bandwidth Reservations under Demand Uncertainty," in Proc. of SIGMETRICS, June 2012.
[13] H. Xu and B. Li, "Maximizing Revenue with Dynamic Cloud Pricing: The Infinite Horizon Case," in Proc. of IEEE ICC, June 2012.
[14] H. Xu and B. Li, "A Study of Pricing for Cloud Resources," ACM SIGMETRICS Performance Evaluation Review, Special Issue on Cloud Computing, March 2013.
[15] W. Wang, B. Li, and B. Liang, "To Reserve or Not to Reserve: Optimal Online Multi-Instance Acquisition in IaaS Clouds," in Proc. of ICAC, June 2013.



Xin Jin received his BEng degree in communication engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2008. He received his Ph.D. degree in Electrical and Electronic Engineering from the University of Hong Kong in 2013. His main research interests are incentive provision, and performance modeling of distributed systems including P2P networks, and cloud computing.

Yu-Kwong Kwok is a Professor in the Electrical and Electronic Engineering Department at the University of Hong Kong (HKU). He received his B.Sc. degree in computer engineering from HKU in 1991, the M.Phil. and Ph.D. degrees in computer science from the Hong Kong University of Science and Technology (HKUST) in 1994 and 1997, respectively. Before joining HKU in August 1998, he was a Visiting Scholar at Purdue University from August 1997 to July 1998. During his sabbatical leave year (from August 2004 to July 2005), he served as a Visiting Associate Professor at the University of Southern California (USC). From 2007 to 2009, he worked as an Associate Professor at the Colorado State University (CSU). He is an Associate Editor for the IEEE Transactions on Parallel and Distributed Systems. He also serves as a member of the Editorial Board for the International Journal of Sensor Networks. From March 2006 to December 2011, he served on the Editorial Board of the Journal of Parallel and Distributed Computing as the Subject Area Editor in Peer-to-Peer (P2P) Computing. He is a Senior Member of the ACM and the IEEE. He is also a member of the IEEE Computer Society and the IEEE Communications Society. He received the Outstanding Young Researcher Award from HKU in November 2004. In January 2010, one of his journal papers was ranked #4 among top ten All-Time Most Cited Papers published in the IEEE Transactions on Parallel and Distributed Systems, based on Scopus and Google Scholar citation counts as of October 2009. In April 2013, he got the Outstanding Reviewer Service Award from the IEEE Computer Society because as of 2013 he was the All-Time Most Prolific Reviewer for the IEEE Transactions on Parallel and Distributed Systems. His recent research endeavors are mainly related to incentive, dependability, and security issues in wireless systems, P2P applications, and clouds.