The Economics of the Cloud: Price Competition and Congestion

JONATHA ANSELMI
Basque Center for Applied Mathematics – BCAM
and
DANILO ARDAGNA
Dip. di Elettronica e Informazione, Politecnico di Milano
and
JOHN C.S. LUI
Computer Science & Engineering, Chinese University of Hong Kong
and
ADAM WIERMAN
Computing and Mathematical Sciences, California Institute of Technology
and
YUNJIAN XU
Computing and Mathematical Sciences, California Institute of Technology
and
ZICHAO YANG
Computer Science & Engineering, Chinese University of Hong Kong

1. INTRODUCTION

The cloud computing marketplace has evolved into a highly complex economic system. Cloud providers are not homogeneous, and a vertical market structure has emerged where *Infrastructure-as-a-Service (IaaS)* and *Provider-as-a-Service (PaaS)* cloud providers rent out the use of (physical or virtual) platforms, servers, storage, networks, etc. While *Software-as-a-Service (SaaS)* deliver applications for users, and often run on top of an IaaS or PaaS. For example, Dropbox is an SaaS running on top of Amazon EC2, which is an IaaS.

Authors' addresses: anselmi@bcamath.org, ardagna@elet.polimi.it, cslui@cse.cuhk.edu.hk, adamw@caltech.edu, xuyunjian@gmail.com, yangtze2301@gmail.com

This letter summarizes our recent work, [Anselmi et al. 2014], which looks at the consequences of this vertical market structure. In particular, we introduce a new model of the cloud marketplace where providers at each layer of the vertical structure are profit maximizing. The key features included by the model are (i) users strategically determine which SaaS provider to use depending on a combination of performance and price; (ii) SaaS providers compete by strategically determining their price and the IaaS/PaaS provider they use in order to maximize profit, which depends on the number of users they attract; (iii) IaaS/PaaS providers compete by strategically determining their price to maximize their profit; (iv) the performance experienced by the users is affected by the congestion of the resources procured at the IaaS/PaaS chosen by the SaaS, and that this congestion is a result of the combination of congestion at *dedicated resources*, where congestion depends only on traffic from the SaaS, and *shared resources*, where congestion depends on the total traffic to the IaaS/PaaS.

Using this model, the goal of our work is to provide insights into fundamental questions such as "How profitable are SaaS providers as compared to PaaS/IaaS providers?" and "Is the market structure such that increased competition among cloud providers yields efficient resource allocation?"

Our analysis highlights a number of qualitative insights with respect to these questions. For example, our results highlight that SaaSs extract profits only as a result of dedicated latency; while IaaS/PaaS providers extract profits from both shared and dedicated latencies. However, the profit of IaaS/PaaS providers reduces significantly as competition grows, and converges to zero in the limit, while services remain profitable even when there are a continuum of services. This highlights that SaaS providers maintain market power over IaaS/PaaS providers even when services are highly competitive, and that one should not expect the cloud marketplace to support a large number of IaaS/PaaS providers.

This observation is similar to the relationship of content providers to ISPs in the internet ([Musacchio et al. 2009; Economides and Tåg 2012]). However, because IaaS/PaaS providers can extract profits from both shared and dedicated latencies they remain reasonably profitable relative to services as long as competition is not extreme. This highlights that the cloud market structure seems not to be as susceptible as the internet to a lack of incentives for infrastructure investment.

But, our analysis highlights an issue with the current market structure: the interaction of SaaS providers and IaaS/PaaS providers serves to protect inefficient IaaS/PaaSs. That is, even if one IaaS/PaaS provider is extremely inefficient compared to another, the inefficient provider still obtains significant profit. Given the suggestion from the results discussed above that the profitability of IaaS/PaaS providers will limit the market to a small level of competition, this "protection" of inefficient providers is a dangerous phenomenon.

2. MODEL OVERVIEW

We begin by briefly summarizing the model introduced in [Anselmi et al. 2014], which focuses on the interaction among three parties in the cloud marketplace: users, SaaS providers (services for short) and IaaS/PaaS providers (providers for short).

Providers: We consider P providers who sell resources to services, as done by Amazon EC2 and Google Cloud. The resources sold can represent virtual machines, in the case of an IaaS, or platforms provided for development, in the case of a PaaS. Each provider p charges a price β_p per unit of data flow for services that use its infrastructure. This charge-per-flow model is very common, e.g., it is used by Google App Engine. We let y_p denote the total flow of provider p and model the profit of provider p by

$$\text{Provider-Profit}(p) = \beta_p y_p. \tag{1}$$

Services: We consider $S \geq 2$ services interacting both with users and providers. Again according to the charge-per-flow model, each service pays the provider that it has chosen to join for infrastructure and charge users for usage. We assume that each service s chooses only one provider, denoted by f_s. So, $f : \{1, \ldots, S\} \to \{1, \ldots, P\}$ is the service-to-provider mapping. Further, each service s charges a unit price α_s to users each time they access to s. Let x_s denote the flow (users/time) of service s, which implies $y_p = \sum_{s:f_s=p} x_s$. Then, the profit of service s is

$$\text{Service-Profit}(s) = (\alpha_s - \beta_{f_s})x_s. \tag{2}$$

Users: The customer base of cloud services is typically quite large, and so we use a nonatomic model in order to capture their aggregate behavior. We model the total user flow to the services as inelastic, and denote it by λ. Thus, we have

$$\lambda = \sum\nolimits_p \sum\nolimits_{s:f_s=p} x_p = \sum\nolimits_p y_p.$$

In the cloud, the latency experienced by users is determined by the combination of both the amount of flow at the service chosen, x_{f_s}, and the amount of flow using the provider chosen by the service y_{f_s}. Thus, we break down the latency experienced into two types of congestion costs: 1) the *dedicated cost (latency)* from the service $\tilde{\ell}_{f_s}(x_{f_s})$ and 2) the *shared cost (latency)* from the provider $\hat{\ell}_{f_s}(y_{f_s})$. Combining these latencies with the service price yields the "effective cost" that users seek to minimize. In particular, for a user who chooses service s, it is modeled by

$$\text{User-Effective-Cost(s)} = \alpha_s + \tilde{\ell}_{f_s}(x_{f_s}) + \hat{\ell}_{f_s}(y_{f_s}). \tag{3}$$

3. EQUILIBRIUM CONCEPTS

To complete the model, we must define the equilibrium concepts used for each level of the model. We give only a high-level overview here, and refer the reader to [Anselmi et al. 2014] for the details.

The key assumption in what follows is that the users act at the fastest time scale, responding to fixed prices of the services and providers, and a fixed mapping of the services to the providers. The next fastest time scale that we consider is pricing, with providers setting prices first and services responding optimally to them. Finally, how services choose the providers to join is modeled as the slowest time scale. This ordering is motivated by the behavior observed in practice: users move quickly between cloud services depending on price, service and provider prices also change quickly (hourly or faster), while the migration of services across providers happens infrequently.

In this context, we first fix how services distribute themselves among providers, and then consider the equilibria of service and provider prices according to a Stackelberg model where providers first set their prices and then services observe these prices and determine the prices they charge to end users. The user flow is then distributed according to a Wardrop equilibrium (cf. the latency cost defined in (3)). The last component to incorporate into our framework is the equilibrium mapping of services to providers, i.e., the distribution equilibrium, which fully characterizes the strategic interaction among the three market participants. At a distribution equilibrium, we have: (i) service and provider prices form a price equilibrium given the particular mapping from services to providers; and (ii) no service has an incentive to change its provider because all providers yield services the same profit.

4. MARKET EFFICIENCY

Given the model described above, one goal of our analysis is to study the market efficiency, as measured through the price of anarchy of user performance. In particular, in [Anselmi et al. 2014] we study the effect of price competition in the cloud on the performance experienced by users using the aggregate user latency resulting from a distribution equilibrium.

Further, to explore the efficiency loss when the number of providers is large, we consider a "replica economy" scaling of providers where there are P types of providers and the number of providers of each type scales with n as n increases to infinity. In this context, as n increases to infinity, we show that there exists an ϵ-equilibrium (among all providers) with ϵ decreasing to zero. We show in [Anselmi et al. 2014] that the price of anarchy of a distribution equilibrium cannot exceed $k+1$, if the latencies are polynomials with degree k.

This result highlights that the price of anarchy will be small in settings when there are a large number of providers. For example, the price of anarchy is bounded by 2 in the case of linear latencies. Interestingly, this is essentially the same price of anarchy as when no market structure exists, i.e., users directly choose providers based on congestion costs [Roughgarden and Tardos 2002]. Since the price of anarchy of the two-tier model (users and SaaSs) converges to one in the limit as the number of services grows [Anselmi et al. 2011], this result reveals that the addition of providers into the marketplace "undoes" the efficiency created by competition among services. Thus, the vertical market structure creates inefficiency that does not exist if SaaSs and PaaSs own their own infrastructure. However, on the positive side, it is this inefficiency that allows the IaaSs to extract profits and avoid the falling prey to the same market failure that doomed ISPs.

5. RELATED WORK

Note that our work in [Anselmi et al. 2014] is a small part of a broader field that focuses on strategic behavior and pricing in cloud systems and, more generally, in the internet.

In the context of cloud systems specifically, an increasing variety of network games have been investigated and three main areas of attention in this literature are resource allocation ([Teng and Magoules 2010; Hong et al. 2011]), load balancing ([Altman et al. 2008; Chen et al. 2009; Anselmi et al. 2011; Anselmi and Gaujal

2011]), and pricing ([Yolken and Bambos 2008; Ardagna et al. 2012; Acemoglu and Ozdaglar 2007; Feng et al. 2013]). It is this last line of work that is most related to the current paper. Within this pricing literature, the most related papers to our work are [Acemoglu and Ozdaglar 2007; Yolken and Bambos 2008; Anselmi et al. 2011; Ardagna et al. 2012; Song et al. 2012; Feng et al. 2013].

Each of these papers focuses on deriving the existence and efficiency (as measured by the price of anarchy) of pricing mechanisms in the cloud. For example, [Ardagna et al. 2012] considers a two-tier model capturing the interaction between SaaSs and a single IaaS, and studies the existence and efficiency of equilibria allocations. Similarly, [Acemoglu and Ozdaglar 2007; Anselmi et al. 2011; Feng et al. 2013] consider two-tier models capturing the interaction between users and SaaSs or between SaaSs and PaaSs/IaaSs, and study the existence and efficiency of equilibrium allocations.

Thus, the questions asked in these (and other) papers are similar to those in our work. However, the model considered in our work is the first to capture the three-tier competing dynamics between users, SaaSs, and IaaSs/PaaSs simultaneously. Further, we model the distinction between congestion from shared and dedicated resources. Neither of these factors was studied in the previous work; and both lead to novel qualitative insights about the cloud marketplace (while simultaneously presenting significant technical challenges to overcome).

REFERENCES

Acemoglu, D. and Ozdaglar, A. 2007. Competition and efficiency in congested markets. *Math. Oper. Res. 32,* 1, 1–31.

Altman, E., Ayesta, U., and Prabhu, B. 2008. Load balancing in processor sharing systems. In *Proc. of ValueTools*. 1–10.

Anselmi, J., Ardagna, D., Liu, J., Wierman, A., Xu, Y., and Yang, Z. 2014. The economics of the cloud: price competition and congestion. *Under submission*.

Anselmi, J., Ayesta, U., and Wierman, A. 2011. Competition yields efficiency in load balancing games. *Perform. Eval. 68*, 986–1001.

Anselmi, J. and Gaujal, B. 2011. The price of forgetting in parallel and non-observable queues. *Perform. Eval. 68,* 12 (Dec.), 1291–1311.

Ardagna, D., Panicucci, B., and Passacantando, M. 2012. Generalized nash equilibria for the service provisioning problem in cloud systems. *IEEE Trans. on Services Computing (Preprint)*.

Chen, H. L., Marden, J. R., and Wierman, A. 2009. On the impact of heterogeneity and back-end scheduling in load-balancing designs. In *Proc. of IEEE INFOCOM*.

Economides, N. and Tåg, J. 2012. Network neutrality on the internet: a two-sided market analysis. *Information Economics and Policy*.

Feng, Y., Li, B., and Li, B. 2013. Price competition in an oligopoly cloud market. *Under submission*.

Hong, Y.-J., Xue, J., and Thottethodi, M. 2011. Dynamic server provisioning to minimize cost in an iaas cloud. In *Proc. of ACM SIGMETRICS*. 147–148.

Musacchio, J., Schwartz, G., and Walrand, J. 2009. A two-sided market analysis of provider investment incentives with an application to the net-neutrality issue. *Review of Network Economics 8,* 1.

Roughgarden, T. and Tardos, E. 2002. How bad is selfish routing? *J. ACM 49*, 236–259.

Song, Y., Zafer, M., and Lee, K.-W. 2012. Optimal bidding in spot instance market. In *Proc. of IEEE INFOCOM*. 190–198.

Teng, F. and Magoules, F. 2010. A new game theoretical resource allocation algorithm for cloud computing. In *Advances in Grid and Pervasive Computing*. 321–330.

Yolken, B. and Bambos, N. 2008. Game based capacity allocation for utility computing environments. In *Proc. of ValueTools*. 1–8.