# The Economics of the Cloud

JONATHA ANSELMI, INRIA Bordeaux Sud-Ouest
DANILO ARDAGNA, Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria
JOHN C. S. LUI, Chinese University of Hong Kong, Department of Computer Science & Engineering
ADAM WIERMAN, California Institute of Technology, Department of Computing & Mathematical Sciences
YUNJIAN XU, Chinese University of Hong Kong, Department of Mechanical & Automation Engineering
ZICHAO YANG, Carnegie Mellon University, Computer Science Department

This article proposes a model to study the interaction of price competition and congestion in the cloud computing marketplace. Specifically, we propose a three-tier market model that captures a marketplace with users purchasing services from Software-as-a-Service (SaaS) providers, which in turn purchase computing resources from either Provider-as-a-Service (PaaS) or Infrastructure-as-a-Service (IaaS) providers. Within each level, we define and characterize market equilibria. Further, we use these characterizations to understand the relative profitability of SaaSs and PaaSs/IaaSs and to understand the impact of price competition on the user experienced performance, that is, the "price of anarchy" of the cloud marketplace. Our results highlight that both of these depend fundamentally on the degree to which congestion results from shared or dedicated resources in the cloud.

CCS Concepts: • **Theory of computation** → **Algorithmic game theory**; **Market equilibria**; *Network games*; • **Computer systems organization** → *Cloud computing*;

Additional Key Words and Phrases: Cloud market, game theory, network economics, equilibrium

**ACM Reference format:**
Jonatha Anselmi, Danilo Ardagna, John C. S. Lui, Adam Wierman, Yunjian Xu, and Zichao Yang. 2017. The Economics of the Cloud. *ACM Trans. Model. Perform. Eval. Comput. Syst.* 2, 4, Article 18 (August 2017), 23 pages.
https://doi.org/10.1145/3086574

## 1 INTRODUCTION

The cloud computing marketplace has evolved into a highly complex economic system made up of a variety of services, which are typically classified into three categories:

This work was supported in part by NSF grants CNS-1518941, CNS-1319820, EPAS-1307794, CNS-0846025, and the Hong Kong RGC GRF 14205114.
Authors' addresses: J. Anselmi; email: jonatha.anselmi@inria.fr; D. Ardagna; email: danilo.ardagna@polimi.it; J. C. S. Lui; email: cslui@cse.cuhk.edu.hk; A. Wierman; email: adamw@caltech.edu; Y. Xu; email: xuyunjian@gmail.com; Z. Yang; email: yangtze2301@gmail.com.

https://doi.org/10.1145/3086574

(1) In *Infrastructure-as-a-Service (IaaS)*, cloud providers rent out the use of (physical or virtual) servers, storage, networks, and so on. To deploy applications, users must install and maintain operating systems, software, and so on. Examples include Amazon EC2, Google Cloud, and Rackspace Cloud.
(2) In *Platform-as-a-Service (PaaS)*, cloud providers deliver a computing platform on which users can develop, deploy, and run their application. Examples include Google App Engine, Amazon Elastic MapReduce, and Microsoft Azure.
(3) In *Software-as-a-Service (SaaS)*, cloud providers deliver a specific application (service) for users. There are a huge variety of SaaS solutions these days, such as email services, calendars, music services, and so on. Examples include services such as Dropbox, Gmail, and Google Docs.

Naturally, each type of cloud service (IaaS, PaaS, SaaS) uses different pricing and contracting structures, which yields a complicated economic marketplace. For example, Amazon computing services are billed on an hourly basis, while some other Amazon services (e.g., queue or datastore) are billed according to the data transfer in and out [AmazonPricing1,AmazonPricing2]. Google App engine pricing is applied on a per-application or user-per-month basis, and more complex billing rules are applied if monthly quotas are exceeded (Google 2014).

Further, adding to the complexity of the cloud marketplace is the fact that a particular SaaS is likely running on top of either a PaaS or IaaS. Thus, there is a multi-tier economic interaction between the PaaS or IaaS and the SaaS and then between the SaaS and the user. This multi-tier interaction was illustrated prominently by the recent crashes of IaaS provider Amazon EC2, which in turn brought down dozens of prominent SaaS providers (Cloudtimes 2012; NetworkWorld 2012).

As a result of the complicated economic marketplace within the cloud, the performance delivered by SaaS providers to consumers depends on both the resource allocation design of the service itself (as traditionally considered) and the strategic incentives resulting from the multi-tiered economic interactions. Importantly, it is impossible to separate these two components in this context. For example, users are both price-sensitive and performance-sensitive when choosing a SaaS; however, the bulk of the performance component for a SaaS comes from the back-end Iaas/PaaS. Further, the IaaS/PaaS does not charge the consumer; it charges the SaaS. Additionally, there is competition among SaaS providers for consumers and among IaaS/PaaS providers for SaaS providers, which yields a competitive marketplace that in turn determines the resource allocation of infrastructure to users and thus the performance experienced by users.

### Contributions of This Paper

This article aims to introduce and analyze a stylized model capturing the multi-tiered interaction between users and cloud providers in a manner that exposes the interplay of congestion, pricing, and performance issues.

To accomplish this, we introduce a novel three-tier model for the cloud computing marketplace. This model, illustrated in Figure 1, considers the strategic interaction between users and SaaS providers (the first and second tiers), in addition to the strategic interaction between SaaS providers and either IaaS or PaaS providers (the second and third tiers). Of course, within each tier there is also competition among users, SaaS providers, and IaaS or PaaS providers, respectively. To the best of our knowledge, this is the first article that jointly considers the interactions and the equilibria arising from the full cloud computing stack (i.e., users, services, and infrastructures/platforms); previous work has focused only on pairwise interactions, for example, Acemoglu and Ozdaglar (2007a), Anselmi et al. (2011), and Ardagna et al. (2012).

The details of the model are provided in Section 2; but, briefly, the key features are as follows: (i) users strategically determine which SaaS provider to use depending on a combination of

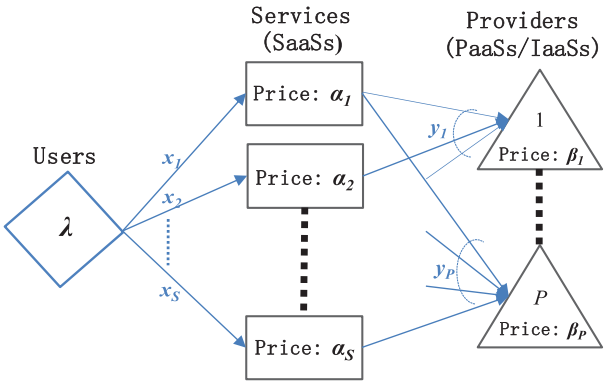


Fig. 1. Overview of model structure and notation.

performance and price; (ii) SaaS providers compete by strategically determining their price and the IaaS/PaaS provider they use 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.

The complex nature of the cloud marketplace means that the model introduced in this article is necessarily complicated, too. To highlight this, note that an analytic study of the model entails characterizing equilibria within each of the three tiers in a context where decisions within one tier impact profits (and thus equilibria) at every other tier.

Due to the complexity of the model, we need to consider a limiting regime to be able to provide analytic results. Motivated by the huge, and growing, number of SaaS providers and the (comparatively) smaller number of IaaS/PaaS providers, the limiting regime we consider is one where the number of users and the number of SaaS providers are both large (see Section 4 for a formal statement). In this setting, we can attain an analytic characterization of the interacting markets that yields interesting qualitative insights.

More specifically, with our analysis, we seek to provide insights into the following fundamental questions:

(1) How profitable are SaaS providers as compared to PaaS/IaaS providers? Does either have market power?
(2) How good is user performance? Is the economic structure such that increased competition among cloud providers yields efficient resource allocation?
(3) How does the degree to which cloud resources are shared/dedicated impact the answers to (i) and (ii)?

Our analysis highlights a number of important, novel qualitative insights with respect to these questions, and we discuss these in detail in Sections 5 and 6. 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 is 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 Internet service providers (ISPs) (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, which is a significant contrast to ISPs. This highlights that the cloud market structure seems not to be as susceptible as the internet to a lack of incentives for infrastructure investment.

Our analysis, on the other hand, 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. The profitability of the inefficient provider is a consequence of *double marginalization* in vertical markets: The efficient provider prefers to charge an extremely high price to match the bad provider's latency cost, because of the existence of profit-maximizing services which also charge prices on users. We note that similar phenomena have been observed in congestion games on parallel-serial networks, where each source-destination path may consist of multiple links operated by independent service providers (Acemoglu and Ozdaglar 2007b). 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.

Another danger that our analysis highlights is that the market structure studied here can yield significant performance loss for users, as compared with optimal resource allocation. We show by example that when providers have highly asymmetric latency cost (e.g., when one provider is much less efficient than the others), an arbitrarily high price of anarchy is possible at an equilibrium due to the double marginalization effect. On the other hand, we prove an upper bound on the price of anarchy that depends only on the minimum and maximum marginal latency costs among all providers. This result provides an efficiency guarantee when all providers are nearly "symmetric" and have similar marginal latency costs. This result highlights that provider symmetry can help to mitigate the potentially high efficiency loss in a cloud computing marketplace.

In an alternative setting, we "fix" the asymmetry of providers by considering a "replica economy" with $P$ types of providers. As the number of providers of each type (and competition) grows, in the limit we show that the price of anarchy cannot be higher than 2, when congestion costs are linear, and $k + 1$ if congestion costs are polynomial with degree $k$. Since the price of anarchy of the two-tier model (users and SaaSs) converges to 1 in the limit as the number of services grows (Anselmi et al. 2011), our result reveals that the addition of providers into the marketplace causes the double marginalization problem and "undoes" the efficiency created by competition among services. Interestingly, this change can be interpreted as inefficiency due to a lack of vertical integration, which would result in two-tier competition. However, vertical integration would lead to other drawbacks, such as making entering the marketplace more challenging, which would reduce participation. Thus, our results highlight that it is crucial to find ways to provide appropriate incentives for the participation of IaaS/PaaS providers in the cloud marketplace, especially given the above observation that profitability of providers decreases quickly with increasing competition.

### Relationship to Prior Work

There is a large literature that focuses on strategic behavior and pricing in cloud systems and, more generally, in the internet. This area of "network economics" or "network games" is full of increasingly rich models incorporating game theoretic tools into more traditional network models. For surveys providing an overview of the modeling and equilibrium concepts in typically used

networking games, and additionally an overview of their applications in telecommunications and wireless networks, see van den Nouweland et al. (1996), Haviv (2001), and Altman et al. (2006).

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 2007a; Feng et al. 2013). It is this last line of work that is most related to the current article. Within this pricing literature, the most related articles to our work are Acemoglu and Ozdaglar (2007a), Yolken and Bambos (2008), Anselmi et al. (2011), Ardagna et al. (2012), Song et al. (2012), Feng et al. (2013), Anselmi et al. (2014), Zheng et al. (2016), and Zhang et al. (2017); see also the references therein.

Each of these articles 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) and Anselmi et al. (2014) consider a two-tier model capturing the interaction between SaaSs and a single IaaS and study the existence and efficiency of equilibria allocations. Similarly, Acemoglu and Ozdaglar (2007a), Anselmi et al. (2011), and Feng et al. (2013) consider again 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) articles are similar to those in our work. However, in contrast to the existing literature, the model considered in this article is the first to capture the three-tier competing dynamics among users, SaaSs, and IaaSs/PaaSs simultaneously. We show that the strategic decisions of any player propagate at all levels, which makes a three-tier model essential to capture the interplay between performance and pricing in the cloud marketplace. 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).

## 2 MODELING FRAMEWORK

We construct a model for studying the interactions among three parities in the cloud marketplace: users, service providers ("services" for short), and infrastructure providers ("providers" for short). In this section, we define the three types of players in our model, but we discuss their strategic behavior only informally. A formal description of the strategic aspects of the model is deferred to Sections 3 and 4. Note that Figure 1 is helpful in understanding the structure of our modeling framework.

**Providers.** We consider $P \geq 2$ 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 $\beta_p$ per unit of data flow for services that use its infrastructure. This charge-per-flow model is very common, for example, 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

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

where, due to the economics of scale in cloud computing, we have ignored the small marginal cost of supporting data flow.

**Services.** We consider $S \geq 2$ services selling the same cloud application (e.g., cloud storage) that interact both with users and providers: They pay infrastructure providers for infrastructure and charge users for usage. We assume that each service $s$ chooses only one (infrastructure) 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 $\alpha_s$ to users. Let $x_s$ denote the flow (users per time unit) 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 therefore we assume a continuum of users having mass $\lambda$, as it is done in nonatomic congestion games. We model the total user flow to the services as inelastic. Therefore, $\lambda = \sum_s x_s$ is constant, where $x$ is the flow vector of services.

When joining a service $s$, users pay $\alpha_s$ to service $s$, as stated above, and incur a congestion cost. In the cloud, congestion is determined by the combination of both the amount of flow at the service chosen, $x_s$, and the amount of flow using the provider chosen by the service $y_{f_s}$. Thus, we further break down the latency experienced into two types of congestion costs: (1) the *dedicated cost (latency)* from the service $\tilde{\ell}_{f_s}(x_s)$ and (2) the *shared cost (latency)* from the provider $\hat{\ell}_{f_s}(y_{f_s})$. The dedicated cost represents congestion cost incurred at the service provider, for example, due to the limited number of virtual machines held by the service. The shared cost represents the congestion at the infrastructure provider, for example, the delay resulting from the network capacity shared by all services using the same infrastructure provider. We note that both the dedicated and the shared latency functions of a service $s$, $\tilde{\ell}_{f_s}(\cdot)$ and $\hat{\ell}_{f_s}(\cdot)$ are determined by its infrastructure provider $f_s$, although the dedicated latency does depend on the user flow of service $s$.

We assume that $\tilde{\ell}_{f_s}(\cdot)$ and $\hat{\ell}_{f_s}(\cdot)$ are continuously differentiable, strictly increasing and convex with $\tilde{\ell}_{f_s}(0) = 0$ and $\hat{\ell}_{f_s}(0) = 0$. Combining these latencies with the service price yields the "effective cost" that users seek to minimize. In particular, the effective cost of a user who chooses service $s$ is

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

In this article, we sometimes focus on linear latency functions, that is, latencies of the form

$$\tilde{\ell}_p(x) = \tilde{a}_p x, \qquad \hat{\ell}_p(y) = \hat{a}_p y, \qquad \forall p, \tag{4}$$

where the slopes $\{\tilde{a}_p\}_{p=1}^P$ and $\{\hat{a}_p\}_{p=1}^P$ are assumed to be positive.

### 2.1 Strategic Interactions and Time-Scale Separation

Throughout this article we interpret the three characters described above as players of a game. Informally, in this game, each provider sets the price that maximizes its individual profit, each service sets the price and chooses the provider that maximizes its individual profit, and each user chooses to join the service that minimizes its individual effective cost.

In practice, these strategic decisions are taken at different time scales. Because of this, it is reasonable to assume that *players acting at a slow time scale will see only the equilibrium behavior of players operating at a faster time scale*. To this end, we assume that the users act at the fastest time scale, responding to fixed prices of the services and a fixed mapping of the services to the providers. The next fastest time scale is service pricing, with services competing with each other to maximize their own profit. Finally, how providers set prices and services decide to distribute among the providers are modeled as the slowest time scale.

Formally, the timing of the game-theoretic model used in this article is as follows. First, there is a simultaneous game among services and providers, where the providers set prices and services choose the connectivity (to providers). Observing the outcome of this simultaneous game, services set their prices for users. Finally, with the information on service prices, users choose their services.

This ordering will be used in the next sections to define strategic equilibria and is motivated by the behavior observed in practice: Users move quickly between cloud services depending on price,

service prices also change quickly (hourly or faster), while the change of provider prices and the migration of services across providers happen infrequently.

## 3 A MODEL WITH ATOMIC SERVICES

In this section, we take a first step toward describing a tractable and reasonable model for the equilibria that result from strategic interactions among users, services, and providers. We consider non-atomic users but atomic services and providers. In this context, we define the equilibria concepts of interest and highlight the analytic difficulties of equilibria characterization. These difficulties motivate the consideration of a model with non-atomic services, which we will introduce in next section and consider in the remainder of the article.

### 3.1 User (Wardrop) Equilibria

Given a fixed service to providers mapping $f$ and fixed service and provider prices, we assume that users distribute to minimize their individual effective cost, defined by Equation (3). Similarly to non-atomic congestion games, for example, Roughgarden and Tardos (2002, 2004), this yields a Wardrop equilibrium (Wardrop 1952), which states that all active services have the same and minimum effective cost. The equilibrium is defined as follows.

*Definition 3.1.* Let mapping $f$ and service prices $(\alpha_1, \ldots, \alpha_S)$ be fixed. A vector $x^{UE} = x^{UE}(\alpha, f) \in [0, \lambda]^S$ is a *user equilibrium* if there exists some $\mu^{UE} \geq 0$ such that

$$\begin{aligned}
&\tilde{\ell}_{f_s}\left(x_s^{UE}\right) + \hat{\ell}_{f_s}\left(y_{f_s}^{UE}\right) + \alpha_s = \mu^{UE}, && \forall s : x_s^{UE} > 0,\\
&\tilde{\ell}_{f_s}\left(x_s^{UE}\right) + \hat{\ell}_{f_s}\left(y_{f_s}^{UE}\right) + \alpha_s \geq \mu^{UE}, && \forall s : x_s^{UE} = 0,\\
&\sum_{s:f_s=p} x_s^{UE} = y_p^{UE}, && \forall p,\\
&\sum_{s} x_s^{UE} = \lambda.
\end{aligned}$$

The existence and uniqueness of a user equilibrium can be easily proven using that conditions in Definition 3.1 coincide with the optimality conditions of a strictly convex optimization problem, as done in Dafermos and Sparrow (1969). This is summarized in the following proposition.

Proposition 3.2. *Let mapping $f$, service prices $\alpha$ and provider prices $\beta = (\beta_1, \ldots, \beta_P)$ be fixed. There exists a unique user equilibrium, which is given by the unique optimal solution of the following strictly convex optimization problem:*

$$\begin{aligned}
\min_{x \geq 0} \quad & \sum_s \left[ \int_0^{x_s} \tilde{\ell}_{f_s}(z)dz + \alpha_s x_s \right] + \sum_p \int_0^{y_p} \hat{\ell}_p(z)dz, \\
\text{s.t.} \quad & \sum_{s:f_s=p} x_s = y_p, \qquad p = 1, \ldots, P, \\
& \sum_s x_s = \lambda.
\end{aligned} \tag{5}$$

### 3.2 Service and Provider Equilibria

We now build on top of the user-level competition described above to consider the price competition of services and providers. In particular, we consider a fixed provider mapping $f$ and define

equilibrium price vectors for services. This is a natural choice for time-scale separation because prices of cloud providers such as Amazon and Google fluctuate minute to minute, but services typically cannot switch between providers at such a fast time scale due to infrastructure setup differences.

In this context, we 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. Of course, the capability to act first confers a strategic advantage for providers over the case where all market participants must choose their moves simultaneously; however, this ordering is natural given the realities of the cloud marketplace.

Given the above discussion, we can now formally define the service equilibrium.

*Definition 3.3.* Let mapping $f$ and provider prices $\beta$ be fixed. A vector $\alpha^{SE} = (\alpha_1^{SE}, \ldots, \alpha_S^{SE})$ is a *service equilibrium* (SE) if

$$\alpha_s^{SE} \in \arg\max_{\alpha_s \geq 0} (\alpha_s - \beta_{f_s})\, x_s^{UE}\left(\alpha_s, \alpha_{-s}^{SE}, f\right), \qquad \forall s. \tag{6}$$

Definition 3.3 is similar to the definition of oligopolistic equilibrium used in Acemoglu and Ozdaglar (2007a) and Hayrapetyan et al. (2007). The essential difference stands in the structure of the latency function of each service $s$, which in our case depends on $x_s$ but also on the whole vector $(x_{s'})_{s':f_{s'}=f_s}$ through $y_{f_s}$. For this reason, existence and uniqueness of a service equilibrium remain difficult issues to study. In fact, as also observed in Acemoglu and Ozdaglar (2007a), service equilibria may not exist if the latency functions are highly convex because this model is a variant of the Bertrand-Edgeworth competition model, which does not admit equilibrium for highly convex cost functions; see Acemoglu and Ozdaglar (2007a) and Maskin and Tirole (1988). The possible non-existence of a service equilibrium makes it difficult to study the interaction among (strategic) providers who set prices to maximize their individual profit, and so it motivates considering a variation of the model for analysis.

As a result, in this article, our approach is to analyze an asymptotic scaling of the model as the number of services becomes large, that is, increases to infinity. See the following section for modelling details and justifications.

## 4 A MODEL WITH NON-ATOMIC SERVICES

The previous section develops the equilibria concepts necessary to characterize the strategic behavior of the users, services, and providers in our model. However, as commented above, there are significant analytic challenges in characterizing these equilibria that make the atomic model intractable to study.

A key observation about cloud markets in practice is that there are generally many more service providers than infrastructure providers. For example, a recent Gartner analysis of the SaaS market have shown how this segment is steadily growing (at the rate of 17.9% (Forbes 2013)), and SaaS cloud market share is already larger than that of IaaS. Moreover, while SaaS started with office suites solutions (e.g., GoogleDocs or Office365), now many vertical market segments have begun to move to the cloud including, for example, ERP (Sap 2014), CRM (Salesforce 2014), Business Process Modeling (Forbes 2013), Project and Portfolio Management, flight scheduling (Ailium 2014), and many others. Even within these individual segments, there is often competition among dozens of firms. As a result, the largest SaaS players (Google and Microsoft) own only 6% of the SaaS market.

Thus, it is natural to consider a situation where there are many more services than providers, while keeping that the number of services is order of magnitude smaller than the number of users. This motivates a change to the model, where services become non-atomic rather than atomic, that is, considering a finite number of providers that can see services as a non-atomic mass. Note that

this is an assumption on the number competing services within a particular market segment, for example, cloud storage where there are more than two dozen competitors, and that this choice parallels the reasoning behind services viewing users as non-atomic. Importantly, the non-atomic model that we define can be interpreted as the limit of the atomic model. Specifically, we show in Appendix A that a symmetric service equilibrium (at which all services that choose the same provider charge the same price) must converge to the non-atomic service equilibrium, as the number of services grows proportionally with the mass of end users and the capacities of the providers. Further, we show by simulation that the non-atomic model is a good approximation for the atomic model even when the number of competing services is small (see Appendix A).

In this section, we introduce the changes to the equilibria concepts that come when non-atomic services are considered. These changes are driven by properties of the atomic model. Importantly, the model becomes much more analytically tractable. In particular, as we show in Sections 5 and 6, for the case with polynomial latency functions, it becomes possible to derive some characterizations for the resulting equilibria that provide interesting insights about market power, profitability, and price of anarchy.

The remainder of this section is organized as follows. In Section 4.1, we first develop a model with non-atomic services that approximates the three-tier market model introduced in the previous section. Then, in Sections 4.2–4.4, we define equilibrium concepts that are based on the non-atomic service model introduced above; proceeding by backward induction to study existence and uniqueness in each case. However, these equilibria are clearly entangled. As a result, the definitions and initial analytic characterizations of the equilibria concepts are intermingled in this section, so the characterizations can aid in simplifying the presentation of the definitions that follow. All proofs are deferred to Appendix B for the ease of the reader.

### 4.1 A Model for Non-atomic Services

We consider a non-atomic model involving a continuum of infinitesimally small services, indexed by $s \in [0, 1]$.

As before, let $\lambda$ denote the total user flow. If the mapping $x_s : [0, 1] \to [0, \infty)$ (from the index of a service to its flow) is Lebesgue measurable, then $\lambda$ can be calculated through the following Lebesgue integral:

$$\lambda = \int_{[0,1]} x_s \mu(ds),$$

where $\mu$ is the Lebesgue measure defined on $[0, 1]$.

Note that, because the latency cost of users depends only on the provider chosen (not the service), all services that choose the same provider are essentially identical. Further, since all services that choose the same provider are faced with the same profit-maximization problem, it is reasonable to assume that they charge the same price to their users.[1] By some abuse of notation, for the rest of the article we will write the price charged by service $s$ as $\alpha_{f_s}$, which depends only on the provider it chooses, $f_s$. Since all users are cost-minimizing, it follows that all services that choose the same provider attract the same amount of data flow, that is, $x_s = x_{s'}$ if $f_s = f_{s'}$. So, for the rest of the article, we use $x_p$ to denote the flow of a service that chooses provider $p$. We can therefore rewrite the profit of a service that chooses provider $p$ as (cf. Equation (2))

$$\text{Service-Profit}(s) = (\alpha_p - \beta_p) x_p, \qquad \forall s : f_s = p.$$

[1]For an atomic model with linear latency functions, we have shown in Appendix A that all services that choose the same provider must set the same price at an equilibrium (cf. Proposition A.2).

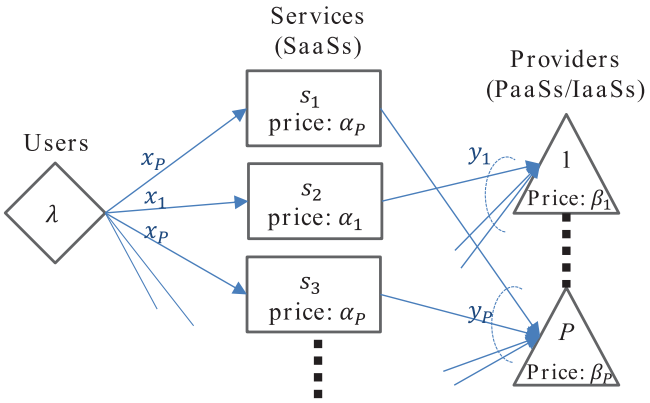


Fig. 2. Overview of the model with non-atomic services.

The non-atomic model introduced in this section is illustrated in Figure 2. Let $g_p$ denote the fraction of services that choose provider $p$ and define a service distribution as a nonnegative $P$-dimensional vector $\mathbf{g} = (g_1, \ldots, g_P)$ such that $\sum_{p=1}^{P} g_p = 1$. Under the assumption that all services associated with a single provider charge their users the same price, we note that different service to provider mappings $f$ that lead to a single service distribution $\mathbf{g}$ will result in the same service prices and user flow. That is, the service to provider mapping $f$ can be fully "represented" by its corresponding service distribution $\mathbf{g}$, and therefore, we will use the latter in the rest of the article. We have

$$y_p = g_p x_p, \qquad \forall p; \qquad\qquad \lambda = \sum_p g_p x_p.$$

### 4.2 User (Wardrop) Equilibrium

Under given service prices $\alpha$ and a service distribution $\mathbf{g}$, user equilibrium can be defined in a way analogous to Definition 3.1.

*Definition 4.1.* Given the prices charged by services $\alpha = (\alpha_1, \ldots, \alpha_P)$ and a service distribution $\mathbf{g} = (g_1, \ldots, g_P)$, a flow vector $\{x_p^{UE}\}_{p=1}^{P}$ is a ***user equilibrium*** if there exists some $\mu^{UE}$ such that

$$\begin{aligned}
&\tilde{\ell}_p\left(x_p^{UE}\right) + \hat{\ell}_p\left(y_p^{UE}\right) + \alpha_p = \mu^{UE}, && \forall p : x_p^{UE} > 0,\\
&\tilde{\ell}_p\left(x_p^{UE}\right) + \hat{\ell}_p\left(y_p^{UE}\right) + \alpha_p \geq \mu^{UE}, && \forall p : x_p^{UE} = 0,\\
&g_p x_p^{UE} = y_p^{UE}, && \forall p,\\
&\sum_p y_p^{UE} = \lambda.
\end{aligned}$$

Further, we denote the set of user equilibria as $W(\alpha, \mathbf{g})$.

Similarly to Proposition 3.2, we have the following characterization on a user equilibrium.

Proposition 4.2. *Given a service price vector $\alpha$ and a service distribution $\mathbf{g}$, there exists a unique user equilibrium, which is the unique optimal solution of the following (strictly) convex optimization*

*problem:*

$$\begin{aligned} \textit{minimize} \quad & \sum_p \left[ \int_0^{x_p} \tilde{\ell}_p(z)dz + \alpha_p x_p \right] + \sum_p \int_0^{y_p} \hat{\ell}_p(z)dz \\ \textit{subject to} \quad & g_p x_p = y_p, \qquad \forall p, \\ & \sum_p y_p = \lambda, \\ & x_p \geq 0, \qquad \forall p. \end{aligned} \tag{7}$$

### 4.3 Service and Provider Equilibria

Before moving to the provider equilibria, let us start with the service (price) equilibrium.

*Definition 4.3.* Given a service distribution $\mathbf{g}$ and a provider price vector $\beta$, a service price vector $\alpha^{SE} = (\alpha_1^{SE}, \ldots, \alpha_P^{SE})$ is a ***service (price) equilibrium***, if

$$\alpha_p^{SE} \in \arg\max_{\alpha \geq 0} (\alpha - \beta_p) x(\alpha, \alpha^{SE}), \qquad \forall p, \tag{8}$$

where

$$\begin{aligned} & x(\alpha, \alpha^{SE}) = 0, && \text{if} \quad \mu^{SE} - \hat{\ell}_p\left(y_p^{SE}\right) < \alpha, \\ & \tilde{\ell}_p(x(\alpha, \alpha^{SE})) + \alpha = \mu^{SE} - \hat{\ell}_p\left(y_p^{SE}\right), && \text{otherwise.} \end{aligned} \tag{9}$$

Here, $y_p^{SE} = g_p x_p^{SE}$, where $(x_1^{SE}, \ldots, x_P^{SE})$ is the unique user equilibrium under the price vector $\alpha^{SE}$ and the service distribution $\mathbf{g}$, and $\mu^{SE}$ is the user effective cost of an active service at the user equilibrium $(x_1^{SE}, \ldots, x_P^{SE})$ (cf. Definition 4.1).

Definition 4.3 is closely related to its atomic counterpart in Definition 3.3. The major difference between these two definitions is that, for an infinitesimally small service that chooses provider $p$, the user equilibrium $(x_1^{SE}, \ldots, x_P^{SE})$ and the corresponding effective cost level $\mu^{SE}$ depend only on the prices set by other services. In Equation (8), $x(\alpha, \alpha^{SE})$ is the user flow attracted by the service, if it sets the price as $\alpha$, and all the other services set their prices according to the equilibrium $\alpha^{SE}$. The value of $x(\alpha, \alpha^{SE})$ is determined by Equation (9). The price $\alpha_p^{SE}$ maximizes the service's profit and yields the service an equilibrium user flow of $x_p^{SE}$, provided that the other services set their prices according to the service equilibrium.

We show in the following proposition that, under a given $\mathbf{g}$ and $\beta$, all service equilibria yield a unique user equilibrium, that is, result in the same user flow. We can therefore use $\mathbf{x}^{SE}(\mathbf{g}, \beta)$ to denote the user equilibrium under a service equilibrium $\alpha^{SE}$ induced by $\beta$ and a service distribution $\mathbf{g}$.

Proposition 4.4. *Given a service distribution $\mathbf{g}$ and a provider price vector $\beta$, there exists a service equilibrium, and all service equilibria result in a unique user equilibrium $\mathbf{x}^{SE}(\mathbf{g}, \beta)$. Further, the equilibrium price of a service who selects a provider $p$ with $x_p^{SE}(\mathbf{g}, \beta) > 0$ is uniquely determined:*

$$\alpha_p^{SE} - \beta_p = x_p^{SE}(\mathbf{g}, \beta) \tilde{\ell}_p'\left(x_p^{SE}(\mathbf{g}, \beta)\right), \qquad \forall p : x_p^{SE}(\mathbf{g}, \beta) > 0. \tag{10}$$

We note that the uniqueness of a user equilibrium under $(\mathbf{g}, \beta)$ together with the service price characterization in Equation (10) implies that all possible service equilibria (induced by a given pair of $\mathbf{g}$ and $\beta$) are the same, except the prices charged by those services with zero user flow (under certain $(\mathbf{g}, \beta)$ it is possible that all services connected to a provider have no user flow). Proposition 4.4 shows that on top of the provider's price $\beta_p$, each service (that attracts positive user flow) earns a per-unit profit that equals the marginal dedicated latency at the induced user equilibrium. In Appendix A, we show that a symmetric equilibrium among atomic services (cf. Definition 3.3) must

converge to the non-atomic service equilibrium characterized in Proposition 4.4, as the number of services increases to infinity.

We are now ready to define the provider (price) equilibrium.

*Definition 4.5.* Given a service distribution $\mathbf{g}$, a provider price vector $\beta^{PE} = (\beta_1^{PE}, \ldots, \beta_P^{PE})$ is a ***provider (price) equilibrium***, if

$$\beta_p^{PE} \in \arg\max_{\beta_p \geq 0} \beta_p x_p^{SE}\left(\beta_p, \beta_{-p}^{PE}, \mathbf{g}\right), \qquad \forall p, \tag{11}$$

where $(x_1^{SE}(\beta_p, \beta_{-p}^{PE}, \mathbf{g}), \ldots, x_P^{SE}(\beta_p, \beta_{-p}^{PE}, \mathbf{g}))$ is the unique user equilibrium induced by provider price vector $(\beta_p, \beta_{-p}^{PE})$ and the service distribution $\mathbf{g}$.[2]

To interpret the above definition, note that, given the prices set by other providers $\beta_{-p}^{PE}$, the equilibrium price $\beta_p^{PE}$ maximizes every provider $p$'s profit at the user equilibrium induced by the price vector $(\beta_p, \beta_{-p}^{PE})$ among all possible prices $\beta_p \geq 0$. In other words, provider (price) equilibrium is a Nash equilibrium among providers under a fixed service distribution $\mathbf{g}$.

Given the definition of provider equilibria, the first questions to address are those of existence and uniqueness. In the case of linear latency functions, we address both these issues by explicitly characterizing the provider price vector $\beta$ at an equilibrium.

Proposition 4.6. *Suppose that latency functions are linear as in Equation (4). Given a service distribution* $\mathbf{g}$ *with at least two positive components,*[3] *there exists a unique provider equilibrium* $\beta^{PE}$ *such that*

*(1) For every p, we have* $x_p^{PE} > 0$, *where* $\mathbf{x}^{PE} \triangleq \mathbf{x}^{SE}(\mathbf{g}, \beta^{PE})$ *is the unique user equilibrium resulting from* $\mathbf{g}$ *and* $\beta^{PE}$ *(cf. Proposition 4.4 for its definition).*

*(2) The provider price vector* $\beta^{PE}$ *is characterized by*

$$\beta_p^{PE} = \left[2\tilde{a}_p + \hat{a}_p g_p + \frac{g_p}{\sum_{p':p' \neq p} \frac{g_{p'}}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}}}\right] x_p^{PE}, \qquad \forall\, p. \tag{12}$$

The above proposition provides a complete characterization on the unique provider equilibrium induced by a service distribution $\mathbf{g}$. In particular, we show that equilibrium provider prices have the same linear structure as the equilibrium service prices characterized in Proposition 4.4, that is, both prices increase linearly with the user flow on the provider/service. Further, it is straightforward to see that an equilibrium provider price increases with the fraction of services it attracts, $g_p$, and decreases with $g_{p'}$ of any other provider $p'$.

Before moving to the next section and introducing our main equilibrium concept, we would like to provide an alternative view of the three equilibrium concepts defined so far, which provides more intuition. In particular, Figure 3 shows an oligopolistic congestion game that mimics both user flow and (service and provider) profit resulting from a provider equilibrium defined above. In this congestion game, each user has to go through two serial links to reach the "destination." An intermediate node represents a provider, and the $p$th node attracts $g_p$ fraction of links (services).

[2]Since $g_p$ is fixed, maximizing the objective function in the definition is equivalent to maximizing its profit $\beta_p x_p^{SE}(\beta_p, \beta_{-p}^{PE}, \mathbf{g}) g_p$. For a provider $p$ with $g_p = 0$, we have implicitly assumed that it aims to maximize the product of its user flow and its unit price, even if the set of services that choose this provider has a zero measure.

[3]If $g_p = 1$ for some provider $p$, then a provider equilibrium does not exist. Since provider $p$ is guaranteed to have a user flow of $\lambda$ (regardless of the price it sets), it would like to charge an arbitrarily high price.

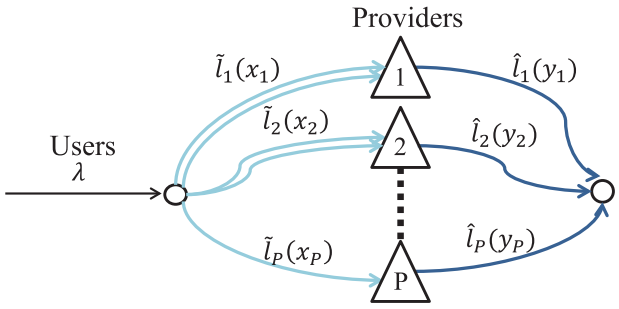


Fig. 3. A congestion game that yields the same user flow (at a Wardrop equilibrium) as that resulting from a provider equilibrium of our model.

The latency of each link is marked in Figure 3, which depends only on the total flow of the link.[4] Once a user chooses its service, its provider $p$ is determined, and the user's cost is given by

$$\tilde{\ell}(x_p) + \hat{\ell}(y_p) + \gamma_p + \beta_p,$$

where $\tilde{\ell}(x_p)$ and $\gamma_p = \alpha_p - \beta_p$ are the latency and price of the first link chosen by the user and $\hat{\ell}(y_p)$ and $\beta_p$ are the latency and price of the second link (that connects the $p$th intermediate node and the destination). We note that the total cost incurred by each user from the source to the destination equals its counterpart in our three-tier model introduced in previous sections. Further, in both the congestion game presented in Figure 3 and our model, the profit of a service that chooses provider $p$ is $\gamma_p x_p$, and provider $p$ obtains a profit of $\beta_p y_p$.

Fixing a service distribution $\mathbf{g}$, we note that the congestion game depicted in Figure 3 has the same payoff structure as our model. As a result, the equilibrium concepts introduced in Definitions 4.3 and 4.5 essentially form a Stackelberg equilibrium of the congestion game where the $P$ links (connected to the destination) simultaneously choose their prices first, and then all the other (non-atomic) links set their prices simultaneously. Proposition 4.4 shows that at all Stackelberg equilibria of this congestion game, each active link in the first segment (corresponding to a service with positive user flow) charges a price $\gamma_p$ that equals its marginal (dedicated) latency at the induced user equilibrium, while the prices charged by the $P$ leaders are characterized in Proposition 4.6.

### 4.4 Distribution Equilibrium

The last component to incorporate into the definition is the mapping of services to providers, that is, the distribution equilibrium, which fully characterizes the strategic interaction among the three market participants. A distribution equilibrium is a triple consisting of service distribution $\mathbf{g}$, provider price vector $\beta$, and service vector $\alpha$ such that

(1) given the service distribution $\mathbf{g}$, the price vectors $(\alpha, \beta)$ form an equilibrium in a Stackelberg game where providers set their prices first;
(2) given the provider price vector $\beta$, the service distribution $\mathbf{g}$ constitutes an equilibrium at which no service can strictly increase its profit by switching to another provider. This is in a spirit similar to that of the well-known Wardrop equilibrium (Wardrop 1952) in that an equilibrium service distribution $\mathbf{g}$ is a steady state evolving after a transient

[4]In contrast to a classical congestion game model, here the total flow of the $p$th link connected to the destination (provider $p$) is $y_p = x_p g_p$.

phase in which profit-maximizing services successively adjust their choices of providers until an equilibrium with stable service profit and user flows have been reached (Larsson and Patriksson 1999). An equilibrium service distribution has an additional level of complication with non-atomic services setting their own prices (while non-atomic users in a Wardrop equilibrium do not).

*Definition 4.7.* A triple, $(\mathbf{g}, \alpha, \beta)$, is a ***distribution equilibrium***, if: (i) $\beta$ is a provider equilibrium under the service distribution $\mathbf{g}$, and $\alpha$ is a service equilibrium under $\mathbf{g}$ and $\beta$ and (ii) no service has an incentive to change its provider because all providers yield services the same profit, that is,

$$x_p^{SE}(\mathbf{g}, \beta)(\alpha_p - \beta_p) = \xi \geq 0, \qquad \forall p : g_p > 0,$$

$$x_p^{SE}(\mathbf{g}, \beta)(\alpha_p - \beta_p) \leq \xi, \qquad \forall p : g_p = 0,$$

where $\mathbf{x}^{SE}(\mathbf{g}, \beta)$ is the unique user equilibrium induced by $\mathbf{g}$ and $\beta$.[5]

Though compact, the above definition tends to be difficult to work with directly. However, in the case of linear latencies, the following conditions are easier to work with and are necessary and sufficient conditions for $(\mathbf{g}, \beta)$ to be a distribution equilibrium, with $\mathbf{x} = (x_1, \ldots, x_P)$ being the user equilibrium resulting from $\mathbf{g}$ and $\beta$ (that is, $\mathbf{x} = \mathbf{x}^{SE}(\mathbf{g}, \beta)$),

$$\begin{cases} \tilde{a}_{p'} x_{p'}^2 = \tilde{a}_p x_p^2, & \text{if} \quad g_p g_{p'} > 0, \qquad (13) \\ \tilde{a}_p x_p^2 \leq \tilde{a}_{p'} x_{p'}^2, & \text{if} \quad g_p = 0, \quad g_{p'} > 0, \qquad (14) \\ 2\tilde{a}_{p'} x_{p'} + \hat{a}_{p'} g_{p'} x_{p'} + \beta_{p'} = 2\tilde{a}_p x_p + \hat{a}_p g_p x_p + \beta_p, & \forall\, p, p', \qquad (15) \\ \beta_p = \left[ 2\tilde{a}_p + \hat{a}_p g_p + \dfrac{g_p}{\sum_{p' \neq p} \frac{g_{p'}}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}}} \right] x_p, & \qquad (16) \\ \sum_p g_p x_p = \lambda, & \qquad (17) \\ \sum_p g_p = 1, & \qquad (18) \end{cases}$$

where (13) and (14) follow from the definition of a distribution equilibrium and the service equilibrium prices characterized in Proposition 4.4. Equation (15) states that at the user equilibrium $\mathbf{x}$, all users have the same effective cost; this is true because all providers have positive user flows (cf. Proposition 4.6). The equality in Equation (16) is the provider equilibrium price[6] characterized in Proposition 4.6. In Section 5, we will discuss the distribution equilibrium in more detail by providing a few illustrative numerical examples.

We can further massage the conditions above to highlight that the distribution equilibrium can be interpreted as a generalized Wardrop equilibrium. In particular, for a triple $(\mathbf{g}, \beta, \mathbf{x}')$ that satisfies the conditions in Equations (13) to (18), we define an alternative user flow vector $\mathbf{x}$ as follows. For any $p'$ with $g_{p'} > 0$, we let $x_{p'} = x'_{p'}$, and for every $p$ with $g_p = 0$, we make $x_p > x'_p$ such that

$$\tilde{a}_p x_p^2 = \tilde{a}_{p'} x_{p'}^2,$$

[5]Note that $\mathbf{x}^{SE}(\mathbf{g}, \beta)$ is defined as the unique user equilibrium resulting from a service equilibrium under $\mathbf{g}$ and $\beta$. Since $\alpha$ is such a service equilibrium, $\mathbf{x}^{SE}(\mathbf{g}, \beta)$ is the unique user equilibrium in $W(\mathbf{g}, \alpha)$.

[6]In the proof of Proposition 4.6, we show that a provider price vector of the form of Equation (16) must be a provider equilibrium (cf. the discussion following Equation (46)).

where $p'$ is a provider with $g_{p'} > 0$. The modified triple $(\mathbf{g}, \beta, \mathbf{x})$ must satisfy the following conditions:

$$\begin{cases} \tilde{a}_{p'} x_{p'}^2 = \tilde{a}_p x_p^2, & \forall p,\ p', & (19) \\ 2\tilde{a}_{p'} x_{p'} + \hat{a}_{p'} g_{p'} x_{p'} + \beta_{p'} = 2\tilde{a}_p x_p + \hat{a}_p g_p x_p + \beta_p, & \text{if} \quad g_p g_{p'} > 0, & (20) \\ 2\tilde{a}_p x_p + \hat{a}_p g_p x_p + \beta_p \geq 2\tilde{a}_{p'} x_{p'} + \hat{a}_{p'} g_{p'} x_{p'} + \beta_{p'}, & \text{if} \quad g_p = 0, \quad g_{p'} > 0, & (21) \\ \beta_p = \left[ 2\tilde{a}_p + \hat{a}_p g_p + \dfrac{g_p}{\sum_{p' \neq p} \frac{g_{p'}}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'}}} \right] x_p, & & (22) \\ \displaystyle\sum_p g_p x_p = \lambda, & & (23) \\ \displaystyle\sum_p g_p = 1. & & (24) \end{cases}$$

Note that, for any triple $(\mathbf{g}, \beta, \mathbf{x})$ that satisfies the preceding conditions (19)–(24), we can construct a triple $(\mathbf{g}, \beta, \mathbf{x}')$ that satisfies the conditions in Equations (13) to (18). It follows that a vector $(\mathbf{g}, \beta)$ that satisfies the conditions in Equations (19)–(24) must be a distribution equilibrium. Next, we define

$$f_p(g_p, g_{-p}) = \frac{1}{\sqrt{\tilde{a}_p}} \left( 2(2\tilde{a}_p + \hat{a}_p g_p) + \frac{g_p}{\sum_{p' \neq p} \frac{g_{p'}}{2\tilde{a}_{p'} + \hat{a}_{p'} g_{p'} S}} \right). \tag{25}$$

Using the above, for a triple $(\mathbf{g}, \beta, \mathbf{x})$ that satisfies the conditions in Equations (19)–(24), substituting $x_p$ and $\beta_p$ to Equations (20) and (21), we obtain

$$\begin{aligned} & f_p(g_p, g_{-p}) = f_{p'}(g_{p'}, g_{-p'}), && \text{if} \quad g_{p'} g_p > 0, \\ & f_p(0, g_{-p}) \geq f_{p'}(g_{p'}, g_{-p'}), && \text{if} \quad g_p = 0, \quad g_{p'} > 0, \\ & \sum_p g_p = 1. \end{aligned} \tag{26}$$

That is, a distribution equilibrium $\mathbf{g}$ must satisfy conditions (26). On the other hand, given a vector $\mathbf{g}$ that satisfies conditions (26), one can solve the triple $(\mathbf{g}, \beta, \mathbf{x})$ by using conditions (19)–(24) and then obtain the corresponding distribution equilibrium $(\mathbf{g}, \beta, \mathbf{x}')$. We conclude that condition (26) is necessary and sufficient for a vector $\mathbf{g}$ to be a distribution equilibrium.

It is this form that provides intuition for the distribution equilibrium concept. In particular, *we can regard condition* (26) *as a generalized Wardrop equilibrium, where the latency function depends on all the components of the model.* Further, it can be verified that $f_p(g_p, g_{-p})$ is a convex function of $\mathbf{g}$ and increases with $g_p$ and decreases with every component of $g_{-p}$. This highlights that, though distribution equilibria are complicated concepts, there is intuition that serves as a guide for our analysis in the coming sections.

## 5 PROFITABILITY

Given the model introduced in the previous two sections, we are now ready to study the interaction of congestion and pricing in the cloud marketplace. The first question we seek to address is the following: Do the providers or services have market power, that is, which extracts the most profit? Then, in the next section, we study the impact of the cloud marketplace on the user experience.

Studying the relative profitability of services and providers requires contrasting the profits attained by services and providers at a distribution equilibrium. However, it is difficult to calculate closed form expressions that allow such a comparison for the general setting. Hence, we consider two special cases of the model here: the case of $P$ symmetric servers and the case of two asymmetric servers. For both cases, we assume that all providers have linear latency functions, so we can obtain simple, interpretable expressions for the service and provider profits.

### 5.1 Symmetric Providers, $P$ Providers

The first case we consider is a symmetric model where $\tilde{a}_p = \tilde{a}$ and $\hat{a}_p = \hat{a}$ for every $p$. In this case, it is easy to characterize the service and provider profits. Specifically, it follows from conditions (26) that there exists a symmetric distribution equilibrium such that

$$g_p = \frac{1}{P}, \qquad p = 1, \ldots, P.$$

Then, from Equation (22), we have

$$\beta_p^{PE} = \frac{P}{P-1}\lambda\left(2\tilde{a} + \hat{a}\frac{1}{P}\right).$$

Through a simple calculation (based on conditions in Equations (13) to (18)), we obtain

$$\text{Provider-Profit}(p) = \left(\frac{2\tilde{a} + \hat{a}/P}{P-1}\right)\lambda^2, \qquad \text{Service-Profit(s)} = \tilde{a}\lambda^2.$$

These expressions for the provider and service profits are quite informative. In particular, they highlight that services extract profits only as a result of dedicated latency in this setting, while providers extract profits from both shared and dedicated latencies. However, competition among symmetric providers significantly reduces the profits providers can extract. Interestingly, competition more quickly reduces the profits that can be extracted from shared latencies than from dedicated latencies. However, as $P \to \infty$, provider profit goes to zero. In contrast, despite the fact that a continuum of services is considered, services still extract positive profit from the marketplace. This highlights that services maintain market power over providers even when services are highly competitive and that one should not expect the cloud marketplace to support a large number of providers.

### 5.2 Asymmetric Providers, 2 Providers

The asymmetric case is more difficult to characterize explicitly, and so we are limited to the case of two providers, $P = 2$. In this setting, we can prove the following proposition, which is the key to our study.

Proposition 5.1. *Consider a case where there are two providers ($P = 2$) with linear latency functions as in Equation (4). There exists a unique distribution equilibrium, which is a solution to the following optimization problem:*

$$\begin{aligned}\text{minimize} \quad & \frac{1}{\sqrt{\tilde{a}_1}}\left(2\left(2\tilde{a}_1 g_1 + \frac{1}{2}\hat{a}_1 g_1^2\right) + 2\tilde{a}_2(-g_1 - \ln(1-g_1)) + \frac{1}{2}\hat{a}_2 g_1^2\right) \\ & + \frac{1}{\sqrt{\tilde{a}_2}}\left(2\left(2\tilde{a}_2 g_2 + \frac{1}{2}\hat{a}_2 g_2^2\right) + 2\tilde{a}_1(-g_2 - \ln(1-g_2)) + \frac{1}{2}\hat{a}_1 g_2^2\right) \\ \text{subject to} \quad & g_1 + g_2 = 1, \quad g_1 \geq 0, \quad g_2 \geq 0.\end{aligned}$$

The proof of the preceding proposition is given in Appendix C.1. Using the preceding proposition, we can explicit calculations comparing the profitability of services and providers in two

(extreme) examples. In particular, we consider examples where one provider is extremely inefficient with respect to either dedicated or shared latencies. These two examples highlight that the competition in the cloud marketplace "protects" inefficient providers. That is, the inefficient provider in both examples still achieves profits within a factor of four of the efficient provider.

### Example: One Provider Has Extremely Inefficient Shared Latency

Consider a setting where $\tilde{a}_1$, $\hat{a}_2$, and $\tilde{a}_2$ are fixed, but the marginal shared cost of provider 1 increases to infinity, that is, $\hat{a}_1 \to \infty$. At a distribution equilibrium, it follows from the characterization of a distribution equilibrium (provided in the proof of Proposition 5.1) that

$$\frac{1}{\sqrt{\tilde{a}_1}}\left(2(2\tilde{a}_1 + \hat{a}_1 g_1) + \frac{2\tilde{a}_2 g_1}{1-g_1} + \hat{a}_2 g_1\right) = \frac{1}{\sqrt{\tilde{a}_2}}\left(2(2\tilde{a}_2 + \hat{a}_2 g_2) + \frac{2\tilde{a}_1 g_2}{1-g_2} + \hat{a}_1 g_2\right).$$

As $\hat{a}_1$ approaches infinity, through a simple calculation, we obtain

$$g_1 \to \frac{\sqrt{\tilde{a}_1}}{\sqrt{\tilde{a}_1} + 2\sqrt{\tilde{a}_2}}, \qquad g_2 \to \frac{2\sqrt{\tilde{a}_2}}{\sqrt{\tilde{a}_1} + 2\sqrt{\tilde{a}_2}}.$$

We then have

$$x_1 \to \frac{\lambda}{3g_1}, \qquad x_2 \to \frac{2\lambda}{3g_2}, \qquad \text{service-profit} \to \frac{\lambda^2}{9}(\sqrt{\tilde{a}_1} + 2\sqrt{\tilde{a}_2})^2,$$
$$\text{provider-profit(1)} \sim \frac{\lambda^2}{9}\hat{a}_1, \qquad \text{provider-profit(2)} \sim \frac{4\lambda^2}{9}\hat{a}_1,$$

where $x_p$ is the equilibrium user flow at provider $p$.

Note that both providers' profits depend only on provider 1's marginal shared cost $\hat{a}_1$ and that the "bad" provider 1 still obtains one half of the user flow of provider 2 and one fourth of the profit of provider 2 despite providing much worse performance.

### Example: One Provider Has Extremely Inefficient Dedicated Latency

Consider a setting where $\hat{a}_1$, $\hat{a}_2$, and $\tilde{a}_2$ are fixed, but the marginal dedicated cost of provider 1 increases to infinity, that is, $\tilde{a}_1 \to \infty$. At a distribution equilibrium, it follows from the proof of Proposition 5.1 that

$$\frac{1}{\sqrt{\tilde{a}_1}}\left(2(2\tilde{a}_1 + \hat{a}_1 g_1) + \frac{2\tilde{a}_2 g_1}{1-g_1} + \hat{a}_2 g_1\right) = \frac{1}{\sqrt{\tilde{a}_2}}\left(2(2\tilde{a}_2 + \hat{a}_2 g_2) + \frac{2\tilde{a}_1 g_2}{1-g_2} + \hat{a}_1 g_2\right).$$

As $\tilde{a}_1$ increases to infinity, through a simple calculation, we have

$$x_1 g_1 \to \frac{\lambda}{3}, \qquad x_2 g_2 \to \frac{2\lambda}{3}, \qquad \text{service-profit} \sim \frac{\lambda^2}{9}\tilde{a}_1,$$
$$\text{provider-profit(1)} \sim \frac{2\lambda^2}{9}\tilde{a}_1, \qquad \text{provider-profit(2)} \sim \frac{8\lambda^2}{9}\tilde{a}_1,$$

where $x_p$ is the equilibrium user flow at provider $p$.

Note that both providers' profit depends only on provider 1's marginal dedicated cost $\tilde{a}_1$ and that, again, the "bad" provider (provider 1) still obtains half of the traffic of provider 2 and one fourth of total profit of provider 2.

## 6 PRICE OF ANARCHY

The second question we study about the cloud marketplace is the following: What is the effect of price competition in the cloud on the performance experienced by users?

To study this question, we measure the "performance experienced by users" by the aggregate user latency resulting from a distribution equilibrium $(\mathbf{g}, \alpha, \beta)$; that is,

$$\ell(\mathbf{x}, \mathbf{g}) \triangleq \sum_p g_p x_p (\tilde{\ell}_p(x_p) + \hat{\ell}_p(g_p x_p)), \tag{27}$$

where $\mathbf{x} = (x_1, \ldots, x_P)$ is the user equilibrium under $\mathbf{g}$ and $\alpha$.

To provide a baseline for comparison, we contrast the aggregate user latency at a distribution equilibrium with the optimal aggregate user latency. That is, we study the "price of anarchy," which is typically used to measure the loss of social welfare caused by the strategic behavior of market participants. In a similar spirit, we define the **price of anarchy (PoA)** of a distribution equilibrium as the ratio of its resulting aggregate user latency to the minimum possible[7]:

$$PoA \triangleq \frac{\ell(\mathbf{x}, \mathbf{g})}{\ell(\mathbf{x}^*, \mathbf{g}^*)}, \tag{28}$$

where $(\mathbf{x}^*, \mathbf{g}^*)$ is an optimal solution to the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \ell(\mathbf{x}, \mathbf{g}), \\ \text{subject to} \quad & \sum_p g_p x_p = \lambda, \\ & \sum_p g_p = 1, \\ & g_p \geq 0, \qquad x_p \geq 0, \qquad \forall p. \end{aligned} \tag{29}$$

Note that a triple $(\mathbf{g}, \beta, \mathbf{x})$ that satisfies conditions (19)–(24) yields the same aggregate latency cost as the corresponding distribution equilibrium $(\mathbf{g}, \beta, \mathbf{x}')$ that satisfies conditions (13) to (18), because the $p$th component of $\mathbf{x}$ is the same as that of $\mathbf{x}'$ for every $p$ with $g_p > 0$. We therefore can (and will) use conditions (19)–(24) to analyze the efficiency of a distribution equilibrium.

The goal of this section is to bound the price of anarchy of the cloud marketplace; however, bounding the price of anarchy in our model under general assumptions is difficult. Thus, throughout this section, we assume that latency functions are polynomial, that is, $\tilde{\ell}_p(x) = \tilde{a}_p x^k$ and $\hat{\ell}_p(y) = \hat{a}_p y^k$, for every $p$.

Under these assumptions, we provide two main results. First, in Section 6.1, we consider a general market model with $P$ providers and we show by example that when one of the providers has very bad latency cost, a distribution equilibrium may yield an arbitrarily high price of anarchy. On the other hand, we prove an upper bound on the price of anarchy that depends only on the minimum and maximum marginal latency costs among all providers. This result provides an efficiency guarantee for a distribution equilibrium, when all providers are nearly "symmetric."

Second, in Section 6.2 we consider an alternative formulation of the model that allows us to separate the impacts of the number of providers and the asymmetry among them. In particular, 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 a sequence of integers $n$ as $n$ increases to infinity.[8]

[7] We note that the aggregate welfare of the system depends only on the aggregate user latency, since the aggregate profit of IaaS and PaaS equals the users' total payment.

[8] Such replica economies are studied commonly in the economics literature, for example, in the context of core convergence (Hart 1979).

In this context, as $n$ increases to infinity, we show that there exists an $\epsilon$-equilibrium with $\epsilon > 0$ decreasing to zero. Further, in the limiting game the price of anarchy is bounded by $k + 1$, which highlights that if the asymmetry of providers is "fixed," then competition among providers leads to efficient performance for users.

### 6.1 General Bounds on the Price of Anarchy

As mentioned above, without any assumptions on the latency cost or the symmetry of the providers, the price of anarchy of the cloud marketplace can be quite large, as highlighted by the following examples.

#### Example: Unbounded Price of Anarchy

Consider a model with $P$ providers. Provider $2, \ldots, P$ are identical, and each has very large marginal shared cost (i.e., $\hat{a}_2 = \hat{a}_3 = \cdots = \hat{a}_P \to \infty$). It is socially optimal for all users to use provider 1, and the minimum aggregate latency cost is given by

$$\ell(\mathbf{x}^*, \mathbf{g}^*) = \sum_p g_p x_p (\tilde{a}_p x_p + \hat{a}_p g_p x_p) = \tilde{a}_1 \lambda^2 + \hat{a}_1 \lambda^2. \tag{30}$$

At a distribution equilibrium, through a simple calculation we obtain

$$g_1 = \frac{2\sqrt{\tilde{a}_1}}{(2\sqrt{\tilde{a}_1} + \sqrt{\tilde{a}_2})}, \qquad g_p = \frac{\sqrt{\tilde{a}_p}}{(P-1)(2\sqrt{\tilde{a}_1} + \sqrt{\tilde{a}_p})}, \qquad p = 2, \ldots, P,$$

which yields a user flow of

$$x_1 = \frac{2\lambda}{3g_1}, \qquad x_p = \frac{\lambda}{(P-1)3g_p}, \qquad p = 2, \ldots, P. \tag{31}$$

It is easy to see that as the marginal shared cost of the $P - 1$ providers increases to infinity, this distribution equilibrium yields an arbitrarily high price of anarchy.

#### Price of Anarchy Bounds for Nearly Symmetric Providers with Polynomial Costs

The previous example highlights that the efficiency of the cloud marketplace depends heavily on the difference between the best and worst providers. This observation leads to the following proposition, which shows that a distribution equilibrium cannot be too inefficient if the worst provider is not "very" different from the best one when the latency cost is polynomial.

Proposition 6.1. *Suppose that latency functions are polynomial, that is,* $\tilde{\ell}_p(x) = \tilde{a}_p x^k$ *and* $\hat{\ell}_p(y) = \hat{a}_p y^k$ *for every* $p$. *The price of anarchy of a distribution equilibrium cannot be higher than*

$$\frac{\tilde{a}_{\max} + \hat{a}_{\max}}{\tilde{a}_{\min} + \hat{a}_{\min}/P^k}, \tag{32}$$

*where* $\tilde{a}_{\min} = \min_p \tilde{a}_p$, $\hat{a}_{\min} = \min_p \hat{a}_p$, $\tilde{a}_{\max} = \max_p \tilde{a}_p$, *and* $\hat{a}_{\max} = \max_p \hat{a}_p$.

Proposition 6.1 is proved in Appendix D.1. It highlights that symmetry of providers is crucial for ensuring the efficiency of the cloud marketplace. The proposition implies that when all providers are symmetric, the price of anarchy converges to one (asymptotic efficiency is achieved) as the number of providers $P$ grows large. Further, it highlights that when the number of providers is large, the ratio of dedicated latency costs to shared latency costs, that is, $\hat{a}_{\max}/\tilde{a}_{\min}$, also plays a significant role in the efficiency of the marketplace.

### 6.2 Bounds on the Price of Anarchy When the Number of Providers Is Large

In this subsection, we consider an alternative formulation of the model that allows us to attain more general bounds on the price of anarchy of the cloud marketplace. In particular, we consider a setting with a large number of small (non-atomic) providers. More specifically, when there are a large number of small providers, it is reasonable to expect that providers cannot anticipate the impacts of their prices on user flow, due to, for example, the lack of information or the limit of computational capability. This assumption leads us to a "non-atomic" provider equilibrium concept for this scenario, which we define below. Then, in this new model, we are able to obtain general bounds on the price of anarchy under polynomial latency cost functions.

*6.2.1 Non-atomic Formulation and Approximation.* In this section, for the case of polynomial latency functions, we define a price equilibrium among a continuum of providers (in the set $[0, 1]$) and show that the non-atomic provider equilibrium serves as a good approximation for its atomic counterpart (cf. Definition 4.5) in a replica economy.

The providers are divided into $P$ types, and for each $p \in [1, \dots, P]$, there is $q_p > 0$ fraction of providers of type $p$. We will focus on symmetric equilibria where all providers of the same type set the same price. As a result, all providers of the same type $p$ must attract the same amount of services, which is denoted by $g_p$ in this section. In this setting, the total amount of services connected to type-$p$ providers is $q_p g_p$, and we have $\sum_p q_p g_p = 1$. Before stating our equilibrium concept, it is useful to specialize some results from previous sections about service equilibrium prices. In particular, it follows from Proposition 4.4 that the service equilibrium prices are

$$\alpha_p^{SE} - \beta_p = x_p \tilde{\ell}_p'(x_p) = k\tilde{a}_p(x_p)^k,$$

which yields users of provider $p$ an effective cost of

$$(x_p)^k \left((k+1)\tilde{\alpha}_p + \hat{\alpha}_p g_p^k\right) + \beta_p.$$

Using the above, we can define the non-atomic provider price equilibrium as follows.

*Definition 6.2.* Given a service distribution $\mathbf{g}$, a provider price vector $\beta^{PE} = (\beta_1^{PE}, \dots, \beta_P^{PE})$ is a ***non-atomic provider (price) equilibrium***, if

$$\beta_p^{PE} \in \arg\max_{\beta \geq 0} \beta x(\beta, \beta^{PE}), \qquad \forall p, \tag{33}$$

where

$$\begin{aligned} &x(\beta, \beta^{PE}) = 0, && \text{if} \quad \mu < \beta, \\ &x(\beta, \beta^{PE})^k((k+1)\tilde{\alpha}_p + \hat{\alpha}_p g_p^k) = \mu - \beta, && \text{otherwise.} \end{aligned} \tag{34}$$

Here, $\mu$ is the user effective cost of an active service at the unique user equilibrium induced by $\mathbf{g}$ and $\beta^{PE}$ (cf. Definition 4.1).

The above definition mimics the definition of a non-atomic service equilibrium in Definition 4.3. Similarly, we have assumed that every provider $p$ is infinitesimally small and therefore has no influence on the user effective cost $\mu$. In Equation (33), $x(\beta, \beta^{PE})$ is the user flow attracted by the provider, if it sets the price as $\beta$, and all the other providers set their prices according to the equilibrium $\beta^{PE}$. The value of $x(\beta, \beta^{PE})$ is determined by Equation (34). The price $\beta_p^{PE}$ maximizes the provider's profit provided that the other providers set their prices according to the equilibrium. It is worth noting that the distribution of provider types $(q_1, \dots, q_P)$ does not show up in the above definition, because each provider's profit depends only on its own price and its user flow, with the latter determined by the service distribution $\mathbf{g}$ and the provider price vector $\beta^{PE}$.

Given the non-atomic provider equilibrium defined above, we can define a corresponding distribution equilibrium that parallels Definition 4.7. The following proposition shows that on top of the provider equilibrium defined in Definition 6.2, both existence and uniqueness of a distribution equilibrium are guaranteed.

Proposition 6.3. *Suppose that latency functions are polynomial, that is,* $\tilde{\ell}_p(x) = \tilde{a}_p x^k$ *and* $\hat{\ell}_p(y) = \hat{a}_p y^k$*, for every p. There exists a unique distribution equilibrium when a nonatomic provider equilibrium is considered.*

The proof of this proposition is given in Appendix D.2. Note that the non-atomic provider equilibrium can be rigorously interpreted as the limit of the original atomic provider game considered to this point of the article. In particular, we justify the non-atomic provider equilibrium concept by considering a replica economy, where there are in total $P$ types of providers, and the number of providers of each type scales with $n$ as $n \to \infty$. More formally, the sequence of finite models defined as follows converges to the model with non-atomic providers we study in this section.

*Definition 6.4.* Consider a sequence of models $\mathcal{G}_n$, $n = 1, 2, \ldots$. For each $\mathcal{G}_n$:

(1) The aggregate user flow is $n\lambda$, and there is a continuum of services in $[0, n]$.
(2) There are a total of $P$ types of providers. The latency functions of a type $p$ provider are assumed to be linear, that is, $\tilde{\ell}_p(x) = \tilde{a}_p x$ and $\hat{\ell}_p(y) = \hat{a}_p y$.
(3) For every $p$, the number of type-$p$ providers is $q_p^n n$, where $\lim_{n\to\infty} q_p^n = q_p$. We assume that $q_p > 0$, for every $p$.

As $n$ increases to infinity, the following proposition shows that every provider's profit is approximately maximized at a distribution equilibrium, as the scaler $n$ (in Definition 6.4) increases to infinity.

Proposition 6.5. *In a sequence of games* $\{\mathcal{G}_n\}_{n=1}^{\infty}$*, a distribution equilibrium among non-atomic providers (on top of the provider equilibrium defined in Definition 6.2) is an* $\epsilon^n$*-equilibrium in the atomic model* $\mathcal{G}_n$*, with* $\epsilon^n$ *decreasing to zero as* $n \to \infty$*.*

The proof of this proposition is given in Appendix D.3.

*6.2.2 Price of Anarchy Results.* Given existence and uniqueness of a distribution equilibrium, we obtain the following bound on the price of anarchy, which is proven in Appendix D.4.

Theorem 6.6. *Suppose that latency functions are polynomial, that is,* $\tilde{\ell}_p(x) = \tilde{a}_p x^k$ *and* $\hat{\ell}_p(y) = \hat{a}_p y^k$*, for every p. The price of anarchy of a distribution equilibrium using a non-atomic provider equilibrium is at most* $k + 1$*.*

In contrast to Proposition 6.1, the above theorem 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 simply 2 in the case of linear latencies, and, more generally, the price of anarchy is $k + 1$ if congestion costs are polynomial with degree $k$. Interestingly, this is essentially the same price of anarchy as when no market structure exists, that is, 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), Theorem 6.2.1 reveals that the addition of providers into the marketplace "undoes" the efficiency created by competition among services. Further, these results highlight that it is crucial to find ways to incentivize participation of IaaS/PaaS providers in the cloud marketplace, especially given the results in Section 5 which highlight that the profitability of providers decreases quickly with increasing competition.

## 7 CONCLUDING REMARKS

In this article, we develop a novel model for the cloud computing marketplace which, for the first time includes (i) the *three-tier* structure of the marketplace (including users, services, and providers) and (ii) the distinction between *shared* and *dedicated* latency in the cloud. The inclusion of these factors leads to novel qualitative insights about market power, user performance (the price of anarchy), and the differing impacts of shared and dedicated latencies.

We view this article as a first step towards a deeper understanding of the cloud marketplace. As such, there are many extensions that are interesting to consider in future work. For example, we have considered one popular price structure, "charge per flow," but there are many other price structures that are available today, including "charge per instance," a fixed "membership" charge, and so on. Additionally, there are many simplifications in the model considered here, for example, that users and services are homogeneous and non-atomic and that there is no market friction preventing services from switching providers. These assumptions are made to allow an analytic first step toward understanding the impact of market structure and would of course be very interesting to remove with future research. A particularly interesting (and challenging) extension to consider would be to study the role of capacity investment decisions of the infrastructure providers. The incorporation of capacity investment decisions into the three-tier model requires considerable new analytic tools but could yield an interesting tradeoff between capacity investment and pricing power.

## REFERENCES


Daron Acemoglu and Asuman Ozdaglar. 2007a. Competition and efficiency in congested markets. *Math. Oper. Res.* 32, 1 (2007), 1–31.

D. Acemoglu and A. Ozdaglar. 2007b. Competition in parallel-serial networks. *IEEE J. Select. Areas Commun.* 25, 6 (2007), 1180–1192.

Ailium. 2014. Flight scheduling as a Service. Retrieved from http://ailium.com.

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

E. Altman, T. Boulogne, R. El-Azouzi, T. Jiménez, and L. Wynter. 2006. A survey on networking games in telecommunications. *Comput. Oper. Res.* 33, 2 (2006), 286–311.

J. Anselmi, D. Ardagna, and M. Passacantando. 2014. Generalized nash equilibria for saas/paas clouds. *Eur. J. Operat. Res.* 236, 1 (2014), 326–339.

Jonatha Anselmi, Urtzi Ayesta, and Adam Wierman. 2011. Competition yields efficiency in load balancing games. *Perform. Eval.* 68 (November 2011), 986–1001.

J. Anselmi and B. Gaujal. 2011. The price of forgetting in parallel and non-observable queues. *Perform. Eval.* 68, 12 (Dec. 2011), 1291–1311. DOI:http://dx.doi.org/10.1016/j.peva.2011.07.023

D. Ardagna, B. Panicucci, and M. Passacantando. 2012. Generalized nash equilibria for the service provisioning problem in cloud systems. *IEEE Trans. Serv. Comput.* 6, 4 (2012), 429–442.

H. L. Chen, J. R. Marden, and A. Wierman. 2009. On the impact of heterogeneity and back-end scheduling in load-balancing designs. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'09)*.

Cloudtimes. 2012. Amazon EC2 outage reveals challenges of cloud computing. Retrieved from http://cloudtimes.org/2012/07/03/amazon-outage-risk-computing/.

S. C. Dafermos and F. T. Sparrow. 1969. The traffic assignment problem for a general network. *J. Res. Natl. Bur. Stand. Ser. B* 73 (1969), 91–118.

N. Economides and J. Tåg. 2012. Network neutrality on the internet: A two-sided market analysis. *Information Economics and Policy* 24, 2 (2012), 91–104.

Y. Feng, B. Li, and B. Li. 2013. Price competition in an oligopoly cloud market. (unpublished).

Forbes. 2013. Gartner Predicts Infrastructure Services Will Accelerate Cloud Computing Growth. Retrieved from http://www.forbes.com/sites/louiscolumbus/2013/02/19/gartner-predicts-infrastructure-services-will-accelerate-cloud-computing-growth/.

Google. 2014. App engine pricing. Retrieved from http://cloud.google.com/pricing/.

O. D. Hart. 1979. Monopolistic competition in a large economy with differentiated commodities. *Rev. Econ. Stud.* 46 (1979), 1–30.

M. Haviv. 2001. The aumann-shapely pricing mechanism for allocating congestion costs. *Operat. Res. Lett.* 29, 5 (2001), 211–215.

A. Hayrapetyan, E. Tardos, and T. Wexler. 2007. A network pricing game for selfish traffic. *Distrib. Comput.* 19 (2007), 255–266.

Yu-Ju Hong, Jiachen Xue, and Mithuna Thottethodi. 2011. Dynamic server provisioning to minimize cost in an IaaS cloud. In *Proceedings of the ACM Conference of the Special Interest Group on Performance Evaluation (SIGMETRICS'11)*. 147–148.

T. Larsson and M. Patriksson. 1999. Side constrained traffic equilibrium models analysis, computation and applications. *Transport. Res. B* 33, 4 (1999), 233–264.

E. Maskin and J. Tirole. 1988. A theory of dynamic oligopoly, II: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica* 56, 3 (May 1988), 571–99.

J. Musacchio, G. Schwartz, and J. Walrand. 2009. A two-sided market analysis of provider investment incentives with an application to the net-neutrality issue. *Rev. Netw. Econ.* 8, 1 (2009), 22–39.

NetworkWorld. 2012. Amazon outage one year later: Are we safer? Retrieved from http://www.networkworld.com/news/2012/042712-amazon-outage-258735.html.

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

T. Roughgarden and E. Tardos. 2004. Bounding the inefficiency of equilibria in nonatomic congestion games. *Games Econ. Behav.* 47 (2004).

Salesforce. 2014. Salesforce Cloud Solutions. Retrieved from http://www.salesforce.com/sales-cloud/overview/.

Sap. 2014. Discover the benefits of cloud computing with SAP. Retrieved from http://www.sap.com/pc/tech/cloud/software/overview/index.html.

Yang Song, Murtaza Zafer, and Kang-Won Lee. 2012. Optimal bidding in spot instance market. In *Proceedins of the IEEE International Conference on Computer Communications (INFOCOM'12)*. 190–198.

F. Teng and F. Magoules. 2010. A new game theoretical resource allocation algorithm for cloud computing. In *International Conference on Grid and Pervasive Computing*. Berlin, Heidelberg, 321–330.

A. van den Nouweland, P. Borm, W. van Golstein Brouwers, R. Bruinderink, and S. Tijs. 1996. A game theoretic approach to problems in telecommunication. *Manage. Sci.* 42, 2 (1996), 294–303.

J. G. Wardrop. 1952. Some theoretical aspects of road traffic research. *Proc. Inst. Civil Eng.* 1 (1952), 325–378.

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

X. Zhang, Z. Huang, C. Wu, Z. Li, and F. Lau. 2017. Online stochastic buy-sell mechanism for VNF chains in the NFV market. *IEEE J. Select. Areas Commun.* 35, 2 (2017), 392–406.

L. Zheng, C. Joe-Wong, C. Brinton, C. Tan, S. Ha, and M. Chiang. 2016. On the viability of a cloud virtual service provider. In *Proceedings of the ACM Conference of the Special Interest Group on Performance Evaluation (SIGMETRICS'16)*. 235–248.

Received March 2017; accepted April 2017