Congestion and price competition in the cloud

Jonatha Anselmi[1], Danilo Ardagna[2], Adam Wierman[3]
[1]BCAM, [2]Politecnico di Milano, [3]California Institute of Technology

***Abstract*—This paper proposes a model to study the interaction of pricing and congestion in the cloud computing marketplace. Specifically, we develop a three-tier market model that captures a marketplace with users purchasing services from Software-as-Service (SaaS) providers, which in turn purchase computing resources from either Provider-as-a-Service (PaaS) providers or Infrastructure-as-a-Service (IaaS) providers. Within each level, we define and characterize competitive 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. Our results highlight that both of these depend fundamentally on the degree to which congestion results from shared or dedicated resources in the cloud. We evaluate the inefficiency of user performance by studying the 'price of anarchy' and show that it grows with the non-linearity of the congestion functions for cloud resources.**

I. Introduction

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

(i) In *Infrastructure-as-a-Service (IaaS)*, cloud providers rent out the use of (physical or virtual) servers, storage, networks, etc. To deploy applications users must install and maintain operating systems, software, etc. Examples include Amazon EC2 and Rackspace Cloud.

(ii) In *Provider-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 and Microsoft Azure.

(iii) 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, ERP, etc. Examples include services such as 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 in the cloud. 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 [5], [6]. Google pricing is applied on a per application or user per month basis and more complex billing rules are applied if monthly quotas are exceeded [15].

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 crash of IaaS provider Amazon EC2, which in turn brought down dozens of prominent SaaS providers [4], [20].

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 paper 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 paper that jointly considers the interactions and the equilibria arising form the players of the full cloud stack (i.e., users, services, and infrastructures/platforms).

The details of the model are provided in Sections II and III but, briefly, the key features 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.

The complex nature of the cloud marketplace means that the model introduced in this paper 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, in order to be able to provide analytic results, we need to consider a limiting regime. 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 III for a formal statement). Under this assumption, we can attain an analytic characterization of the interacting markets which yield interesting qualitative insights.

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

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

Our analysis highlights a number of important qualitative insights with respect to these questions, and we discuss these in detail in Section IV. For example, our analysis shows that if congestion is dominated by the congestion at shared resources, the cloud market does not function well, i.e., providers can be profitable but services are unprofitable and thus have no reason to participate in the market. In contrast, if congestion is dominated by congestion at dedicated resources, markets function well, i.e., providers and services are both profitable and services receive the bulk of the profits. However, in both cases, our analysis highlights another issue with the current market structure: the interaction of service and provider markets serves to protect inefficient providers. That is, even if one provider is extremely inefficient compared to other, the inefficient provider still obtains significant profit. Finally, 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. Specifically, the price competition among services and providers yields inefficient resource allocation.

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' of 'network games' is full of increasingly rich models incorporating game theoretic tools into more traditional network models. The following surveys provide an overview of the modeling and equilibrium concepts in typically used networking games, and additionally include an overview of their applications in telecommunications and wireless networks [3], [17], [25].

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 [18], [24], load balancing [2], [7], [8], [12], and pricing [1], [9], [14], [27]. 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 [1], [7], [9], [14], [23], [27]; see also the references therein. Each of these papers focus on deriving the existence and efficiency (as measured by the price of anarchy) of pricing mechanisms in the cloud. For example, [9] considers a two-tier model capturing the interaction between SaaSs and a single IaaS, and studies the existence and efficiency of equilibria allocations. Similarly, [1], [7], [14] 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 this paper is significantly more general than prior work studying the cloud marketplace. Specifically, we 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).

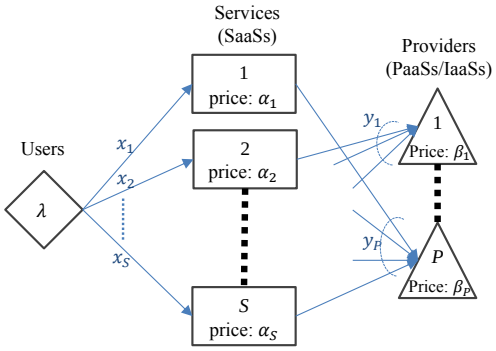

Fig. 1. Model overview.

II. Model Overview and Notation

Our model represents a three-tier cloud marketplace and focuses on the interaction between PaaS or IaaS providers (termed "providers"), SaaS providers (termed "services"), and final customers (termed "users"). We discuss each of the three tiers in the following, starting with providers. The model is detailed, and so Figure 1 provides a high level diagram.

This section focuses on the non-competitive aspects of the model and the next one introduces, and begins to characterize, the competitive equilibria considered at each level.

A. Providers

The "providers" in our model represent the PaaSs and IaaSs in the cloud. That is, they sell computing infrastructure to services (SaaSs) in a manner such as done by Amazon EC2 or Google App Engine.

We model competition among $P \geq 2$ providers, where each provider p sells capacity to services at a price of β_p per unit of capacity and time. For simplicity, we assume throughout that services are all interested in one unit of capacity. This pricing setup is a simplified model of the pricing scheme that is currently employed by Amazon EC2 [5].

The total traffic served by provider p is denoted by y_p. Note that y_p is dependent on both the fraction of services that choose provider p, which we denote by g_p, and the number of users that choose those services. Since g_p is the fraction of services choosing provider p, $g_p S$ is the number of services that use provider p, where $S \geq 2$ denotes the number of services. We use this particular formulation because we are interested in letting the number of services grow large in our analysis.

The profit for provider p, per unit of time, is thus:

$$\text{Provider-Profit}(p) = \beta_p g_p S.$$

Note that we do not model the operating costs of the providers, but this could be incorporated easily into the above equation.

B. Services

The "services" in our model play the role of SaaSs. That is, they sell a service to users, but buy the computing infrastructure used to provide this service from a provider.

We model competition among $S \geq 2$ (interchangeable) services, where each service s is sold for a price α_s to users. The user traffic to service s is denoted by x_s. Each service is run on exactly one of the P providers, and the provider chosen by service s is denoted by f_s. For simplicity, we assume each service is interested in one unit of service capacity that provider p sells at price β_p.

The profit for service s, per unit of time, is thus:

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

Note that we do not model the operating costs of the services, but this could be incorporated easily into the above equation.

C. Users

The "users" in our model play the role of the customers of services (SaaSs). Since these services have enormous user bases, we consider a model with competition among infinitely-many users, each of which controls an infinitesimal amount of traffic, i.e., a non-atomic routing game. In total, the users make up an (inelastic) arrival rate of λ.

Upon arriving to the system, each user selects exactly one service. When making this selection there are two factors that come into play: the price charged by the service $\alpha_s \geq 0$ and the congestion-dependent cost $\ell_s(x, f)$, where $x = (x_1, \ldots, x_S)$ is the amount of traffic going to each service and f is the service-to-provider mapping. This cost can be interpreted as the mean delay or mean reliability of service s. We use the terms congestion cost and latency interchangeably.

It is important to highlight the dependence of the congestion cost on both x and f in the setting of this paper. Specifically, the congestion cost at service s could be affected by both the number of users choosing s and also the number of users choosing other services deployed on the same provider of s, i.e., f_s. This is because providers may have some dedicated resources for services (such as virtual machines, servers, or even a cluster of computers); but other resources at the provider are necessarily shared (such as networking components and possibly storage). In this paper, in order to obtain analytic results, we limit ourselves to a particular form of interaction between shared and dedicated resources:

$$\ell_s(x, f) = \tilde{\ell}_{f_s}(x_s) + \hat{\ell}_{f_s}(y_{f_s}). \tag{1}$$

So, $\tilde{\ell}$ captures the congestion at dedicated resources (congestion depends on only the considered service traffic) and $\hat{\ell}$ captures the congestion at shared resources (congestion depends on the total traffic to the considered provider). Of course, one could analyse generalizations of the form given in (1); however this form already captures the qualitative interaction of these types of workloads. In fact, the relative magnitude of congestion in shared and dedicated resources shows up prominently in the results provided in Section IV.

Some additional technical assumptions (widely used in the literature, e.g., [1], [13]) we require are that $\tilde{\ell}_p : \mathbb{R}^+ \to \mathbb{R}^+$ and $\hat{\ell}_p : \mathbb{R}^+ \to \mathbb{R}^+$, are continuously-differentiable, increasing and convex; $\tilde{\ell}_p(0) = \hat{\ell}_p(0) = 0$ for all p; and $\tilde{\ell}_p(x_s) < \infty$ for all x_s and $\hat{\ell}_p(y_p) < \infty$ for all y_p.

Given the above, we model the users as minimizing their *effective-cost* when choosing a service s:

$$\begin{aligned}\text{User-Effective-Cost}(s) &= \alpha_s + \ell_s(x, f) \\ &= \alpha_s + \tilde{\ell}_{f_s}(x_s) + \hat{\ell}_{f_s}(y_{f_s}).\end{aligned}$$

This form is motivated by considering the congestion cost as being in currency units, and is a common modeling assumption in the literature on congestion games; e.g., [1], [13].

III. Modeling and analyzing competition

A crucial component of our model is capturing the strategic behavior of users, services, and providers. In this section, we define the equilibrium concepts we study within each level. Clearly, these equilibrium concepts are quite entangled. So, we start by defining the user equilibrium and build to the provider equilibrium. Along the way, we derive results characterization and existence results for each of the equilibrium concepts introduced. Throughout, proofs are differed to the appendix for the sake of conciseness.

A. Competition among users

Since each user carries an infinitesimally-small amount of traffic, we model competition among users as a *non-atomic* routing game; e.g., [21]. In this context, effective-cost minimizing competition yields a Wardrop equilibrium, i.e., a traffic allocation that follows from Wardrop principles [26]: the effective cost of each user is identical and minimum at each used service. Using the notation of our model, a user equilibrium is defined as follows.

Definition 1. *Consider a fixed service mapping f and service prices α. A vector $x^{UE} = x^{UE}(\alpha, f) \in [0, \lambda]^S$ is a **user equilibrium** if*

$$\begin{aligned}&\tilde{\ell}_{f_s}(x_s^{UE}) + \hat{\ell}_{f_s}(y_{f_s}^{UE}) + \alpha_s \\ &= \min_{s' : x_{s'}^{UE} > 0} \left\{ \tilde{\ell}_{f_{s'}}(x_{s'}^{UE}) + \hat{\ell}_{f_{s'}}(y_{f_{s'}}^{UE}) + \alpha_{s'} \right\}, \forall s : x_s^{UE} > 0 \\ &\textstyle\sum_{s : f_s = p} x_s^{UE} = y_p^{UE}, \ \forall p, \\ &\textstyle\sum_s x_s^{UE} = \lambda.\end{aligned} \tag{2}$$

This is a well established equilibria concept, e.g., [10], and it is easy to obtain a strong characterization of the user equilibrium using a potential function-based argument showing that conditions (2) are the KKT conditions of a strictly-convex optimization problem.

Proposition 1. *Consider a fixed service mapping f and service prices α. There exists a unique user equilibrium, which is given by the optimizer of*

$$\begin{aligned}\min_{x \geq 0} \quad & \sum_s \left[\int_0^{x_s} \tilde{\ell}_{f_s}(z) dz + \alpha_s x_s \right] + \sum_p \int_0^{y_p} \hat{\ell}_p(z) dz, \\ \text{s.t.:} \quad & \textstyle\sum_{s : f_s = p} x_s = y_p, \ \forall p \\ & \textstyle\sum_s x_s = \lambda.\end{aligned} \tag{3}$$

Importantly, because optimization (3) is strictly convex, the unique user equilibrium is efficiently computable [11].

B. Competition among services

We model competition among services as an oligopolistic pricing game between $S \geq 2$ profit-maximizing services as in [1], [7]. In particular, we assume that each service sets a price and selects a provider in order to maximize its profit, yielding a Nash equilibrium where no service has a unilateral incentive to deviate.

Services have two choices (price and provider), and thus the equilibrium needs to consider both. Because prices may fluctuate at a faster time scale than the choice of a provider, we focus on a two stage equilibrium and first define a "service-equilibrium price vector" before then using this concept to define a "service-equilibrium distribution".

Service-equilibrium price vector: We define a service-equilibrium price vector as follows.

Definition 2. *Consider a fixed service mapping f. Then, $\alpha^{SE} = \alpha^{SE}(f)$ is a **service-equilibrium price vector** if*

$$\alpha_s^{SE} \in \underset{\overline{\alpha}_s \geq 0,}{\operatorname{argmax}}\ \overline{\alpha}_s x_s^{UE}((\overline{\alpha}_s, \alpha_{-s}^{SE}), f),\ \forall s. \tag{4}$$

Establishing the existence of a service-equilibrium price vector is harder than in the case of user equilibria. To highlight this, note that in some special cases, a service-equilibrium price vector is equivalent to the notion of oligopolistic equilibrium in [1], for which existence is shown only when the latency functions are linear.

So, when characterizing service-equilibrium price vectors, we also limit ourselves to linear latency functions. Within this context, the following proposition shows the existence of a service-equilibrium price vectors even when $\hat{\ell}_p(\cdot) \neq 0$, provided that both $\tilde{\ell}_s(\cdot)$ and $\hat{\ell}_p(\cdot)$ are linear. It is likely that existence can be guaranteed outside of these restrictions as well; however, as in [1], [7], there are cases where an equilibrium does not exist.

Proposition 2. *Consider $\tilde{\ell}_{f_s}(x_s) = \tilde{a}_{f_s} x_s$ and $\hat{\ell}_p(y_p) = \hat{a}_p y_p$ for all s and p. There exists a unique service-equilibrium price vector.*

We can also characterize the structure of service-equilibrium price vectors in a more general context, provided they exist.

Proposition 3. *Consider a fixed service mapping f. If α^{SE} is a service-equilibrium price vector, then $\alpha_{s'}^{SE}$ satisfies (5) for all s', where $p' = f_{s'}$ and $x = x^{UE}(\alpha^{SE}, f)$ is a user equilibrium.*

The price structure described by (5) is clearly cumbersome. Given that the service-equilibrium distribution and the provider equilibrium (defined later) both build on this characterization, it is important to find a simpler representation if we hope to be able to obtain analytic results.

The important observation we use to obtain such a simplification is that, in practice, the number of services tends to be very large while the number of providers is (comparably) small. Thus, a setting with $S \gg P$ is reasonable. Further, it happens that (5) simplifies dramatically when we consider a large number of services. These observations lead us to consider the following "large-system limit".

Definition 3. *The **large-system limit** is defined via a scaling parameter $n \in \mathbb{N}$, with $n \to \infty$, where in the nth system:*

(i) The user arrival rate is $n\lambda$.

(ii) There are nS services and where services $s, S+s, 2S+s, \ldots, (n-1)S+s$ choose the same provider, for all $s = 1 \ldots, S$.

(iii) The dedicated and shared resource capacity of each provider scales proportionally with n, i.e., the latency obeys $\ell_s(x) = \tilde{\ell}_{f_s}(x_s) + \hat{\ell}_{f_s}(\frac{y_{f_s}}{n})$, for all s.

This limiting regime corresponds to a situation where our system is 'replicated' n times. This is a classic approach in economics [16] and in our case yields the interpretation that S is the number of 'normalized' services, each of which behaves like a non-atomic player. In this regime, the characterization of the price vector equilibrium becomes manageable.

Proposition 4. *Consider a fixed service mapping f. In the large-system limit, there exists at most one service-equilibrium price vector α^{SE} and*

$$\alpha_s^{SE} \to x_s^{UE} \tilde{\ell}'_{f_s}(x_s^{UE}),\ \forall s, \tag{6}$$

where $x^{UE} = x^{UE}(\alpha^{SE}, f)$ is a user equilibrium.

Not only is the price structure in (6) more manageable than (5), it also highlights some important insights. Specifically, (6) states that, in the large-system limit, services make profit only because of the congestion on dedicated resources, congestion at shared resources is irrelevant. In other words, the negative externalities exerted to users due to the congestion on shared resources is not profitable for services.

Service-equilibrium distribution: We have now defined and characterized the equilibrium with respect to one of the choices that services face, i.e., prices. What remains is to define and characterize the equilibrium for how services distribute over providers, i.e., the service-equilibrium distribution.

Clearly, the service-equilibrium distribution depends on both the user equilibrium and the service-equilibrium price vector. The following lemma highlights that all the services mapped to the same provider have identical service-equilibrium prices and incoming traffic.

Lemma 1. *For any s_1, s_2 such that $f_{s_1} = f_{s_2}$, $\alpha_{s_1}^{SE}(f) = \alpha_{s_2}^{SE}(f)$. Furthermore, $x_s^{UE}(\alpha^{SE}, f) = z_{f_s}^{UE}$ for all s where $z^{UE} \stackrel{\text{def}}{=} (z_1^{UE}, \ldots, z_P^{UE})$ is the unique solution of*

$$\begin{aligned} &\min_{p': z_{p'}^{UE}, g_{p'} > 0} \left\{ \tilde{\ell}_{p'}(z_{p'}^{UE}) + \hat{\ell}_{p'}(g_{p'} S z_{p'}^{UE}) + \alpha_{p'}^{SE} \right\} \\ &= \tilde{\ell}_p(z_p^{UE}) + \hat{\ell}_p(g_p S z_p^{UE}) + \alpha_p^{SE},\ \forall p : z_p^{UE}, g_p > 0 \\ &\textstyle\sum_p g_p S z_p^{UE} = \lambda \\ &z_p^{UE} > 0 \text{ if and only if } g_p > 0,\ \forall p \end{aligned} \tag{7}$$

and $\alpha_p^{SE} = \alpha_p^{SE}(f)$ is a service-equilibrium price of any service that is mapped on provider p, for all p.

The above lemma implies that services are "indistinguishable" from the providers' standpoint and allows us to define a service-equilibrium distribution in terms of g, the proportion of services using each provider, instead of f, the exact mapping of services to providers. This is an important shift since we are focused on the large-system limit. In the following, recall that $g_p = g_p(f) = |\{s : f_s = p\}|/S$.

$$\alpha_{s'}^{SE} = x_{s'}\ell'_{p'}(x_{s'}) + x_{s'}\frac{\frac{\sum_{s:f_s\neq p'}\left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}}{\left(\hat{\ell}'_{p'}(y_{p'})\right)^{-1}-\sum_{s:f_s\neq p'}\left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}}+1}{\frac{\sum_{s:f_s\neq p'}\left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}}{\left(\hat{\ell}'_{p'}(y_{p'})\right)^{-1}-\sum_{s:f_s\neq p'}\left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}}\sum_{s\neq s':f_s=p'}\left(\tilde{\ell}'_{p'}(x_s)\right)^{-1}} + \sum_{p\neq p'}\frac{\left(\sum_{s:f_s=p}\left(\tilde{\ell}'_{p}(x_s)\right)^{-1}\right)^2}{\left(\hat{\ell}'_{p}(y_{p})\right)^{-1}+\sum_{s:f_s=p}\left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}} + \sum_{s\neq s'}\left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}} \tag{5}$$

Definition 4. *Consider fixed provider prices β. Then $g^{SE} = g^{SE}(\beta)$ is a **service-equilibrium distribution** if it solves*

$$\begin{aligned} &\alpha_p^{SE} z_p^{UE} - \beta_p = \max_{p':g_{p'}^{SE}>0}\left\{\alpha_{p'}^{SE} z_{p'}^{UE} - \beta_{p'}\right\},\ \forall p : g_p^{SE} > 0 \\ &\textstyle\sum_p g_p^{SE} = 1 \\ &g_p^{SE} \geq 0,\ \forall p \end{aligned} \tag{8}$$

where $z_p^{UE} = z_p^{UE}(g^{SE})$ satisfies (7)*, and $\alpha_p^{SE} = z_p^{UE}\tilde{\ell}'_p(z_p^{UE})$, i.e., it satisfies* (4)*.*

Note that this definition corresponds to a variation of Wardrop equilibrium [21], [26], similar to the concept used for the user equilibrium. This setting is reasonable since services are indistinguishable and their number is large, so no particular service has measurable market power. Specifically, in this equilibrium the proportions of services on each provider are such that the profits of all services are equal and maximal. We stress that Definition 8 is dependent on the scaling in Definition 5. The term S that appears in the definition of $z^{UE}(g^{SE})$ should be interpreted as the "normalized" number of services. So, Sg_p^{SE} can be any non-negative number.

We now move to characterizing service-equilibria distributions. Proving the existence of a service-equilibrium distribution is complicated by the discontinuities (in g) that emerge in the equations characterizing $z^{UE}(g)$, i.e., (7). The following proposition gives a condition for which the service-equilibrium distribution is unique, if it exists, and can be found via a strictly-convex optimization problem.

Proposition 5. *Consider fixed provider prices β. Let $(g^\star(\beta), y^\star(\beta))$ be the unique optimizer of the following strictly-convex optimization problem*

$$\begin{aligned} \min_{g\geq 0, y\geq 0}\quad & \sum_p y_p\tilde{\ell}_p\left(\frac{y_p}{Sg_p}\right) + \int_0^{y_p}\hat{\ell}_p(z)dz + S\beta_p g_p \\ \textit{s.t.:}\quad & \textstyle\sum_p g_p = 1, \\ & \textstyle\sum_p y_p = \lambda. \end{aligned} \tag{9}$$

If $g_p^{SE}(\beta) > 0$ and $g_p^\star(\beta) > 0$ for all p, then $g^{SE}(\beta) = g^\star(\beta)$ is the unique service-equilibrium distribution.

Under the assumption that $g_p^{SE}(\beta) > 0$ for all p, the optimizers of (9) coincide with the conditions that define a service-equilibrium distribution (Definition 4), which is a remarkable and nontrivial property of our model. It will be shown in Lemma 2 that the assumption $g_p^{SE}(\beta) > 0$ for all p is true if β forms a provider equilibrium (defined next).

C. Competition among providers

To capture competition among providers, we consider price competition among $P \geq 2$ providers that are profit maximizing. This yields the following Nash equilibrium formulation.

Definition 5. *The vector β^{PE} is a **provider equilibrium** if*

$$\beta_p^{PE} \in \operatorname*{argmax}_{\overline{\beta}_p}\ \overline{\beta}_p\, g_p^{SE}(\overline{\beta}_p, \beta_{-p}^{PE})S,\ \forall p. \tag{10}$$

Importantly, this equilibrium embeds the equilibrium at the service level, which in turn captures the equilibrium at the user level. As such, it is the most informative, but also the most difficult to study.

Characterizing the existence of provider equilibria is a very difficult task. This is highlighted by observing the structure of optimization (10), which is non-concave even when latency functions are linear. As a result, we do not provide analytic results regarding existence. However, it is possible to obtain results characterizing provider equilibria, as done in Proposition 3 for service-equilibrium prices in a special case. In particular, the following proposition establishes the structure of the provider equilibrium prices in the case of two providers and linear latency functions. The structure provided by this result (i) provides a constructive numerical method to check existence of a provider equilibrium, as we discuss next, and (ii) is crucial for characterizing the user performance and service/provider profits, as we do in Section IV.

Lemma 2. *If β is a provider equilibrium, then $g_p^{SE}(\beta) > 0$ for all p.*

Proposition 6. *Let $P = 2$, $\tilde{\ell}_{f_s}(x_s) = \tilde{a}_{f_s}x_s$ and $\hat{\ell}_p(y_p) = \hat{a}_p y_p$, for all s, p. Then*

$$\beta_{p'}^{PE} = 2g_{p'}\sum_p \tilde{a}_p\frac{z_p^2}{g_p} - \frac{4g_{p'}}{S}\frac{\left(\sum_p \tilde{a}_p\frac{z_p}{g_p}\right)^2}{\sum_p \hat{a}_p + \frac{2\tilde{a}_p}{Sg_p}},\ p' = 1, 2 \tag{11}$$

where $g = g^{SE}(\beta^{PE})$ is a service-equilibrium distribution and $z = z^{UE}(g)$ satisfies (7)*.*

An important consequence of Proposition 6 is that it allows numerical investigation of provider equilibria. In particular, upon substituting (11) into conditions (8), one obtains a nonlinear system composed of four equations and four unknowns, i.e., the y_p and g_p. Substituting the solutions of this system back in (11), one obtains a guess for a provider equilibrium. To check whether or not this guess (say $\overline{\beta}$) is a provider equilibrium, one can check numerically (10), i.e., whether or not $\overline{\beta}_p$ is the best-response to $\overline{\beta}_{-p}$. Using this approach we have numerically studied 100 random instances which $\hat{a}$ and $\tilde{a}$ chosen uniformly at random from [0.1,10]. In all cases, we found a unique solution to this system with the resulting provider prices forming a best-response, which verifies that provider equilibria typically exist when latencies are linear.

IV. Profitability and efficiency

To this point, we have introduced a model for the cloud computing marketplace, and characterized the equilibria that

result from competition within this setting. Our goal in this section is to use this characterization to understand the impact of these interacting markets on all the parties involved, i.e., the users, services, and providers. In particular, our goal is to study the three questions outlined in the Introduction.

Note that the analysis presented in the following crucially uses Propositions 5 and 6, which characterize the service and provider equilibria respectively.

A. Profitability of services and providers

In order to study the relative profitability of services and providers, we need to make use of a detailed characterization of provider equilibria, and so our focus is on the special case of $P = 2$ and linear latency functions, as this restriction is necessary to apply Proposition 6.

To apply Proposition 6, we first need to compute the user equilibrium traffic allocation and the service equilibrium price vector and distribution. These are characterized by conditions (8) where the β_p are given by (6). One can show that the resulting system of equations can be expressed in terms of the solution a polynomial of order four. Thus, finding the solutions of this system is possible both analytically and numerically. However, the resulting expression for g^{SE} is cumbersome and thus omitted. Instead, we focus on representative special cases and numeric examples in order to highlight the insights one can learn from the expression. In particular, in the following we focus on the cases of symmetric and asymmetric latencies.

Symmetric latencies: To begin, we consider the special case when latency functions are symmetric. In this setting, we obtain simple, informative characterizations of the provider equilibrium and service equilibrium price vector.

Corollary 1. *Let $P = 2$, $\tilde{\ell}_{f_s}(x_s) = \tilde{a}x_s$ and $\hat{\ell}_p(y_p) = \hat{a}y_p$, for all s, p. Then, $g^{SE}(\beta^{PE}) = (\frac{1}{2}, \frac{1}{2})$, $\alpha^{SE}(f) = \tilde{a}\frac{\lambda}{S}$, with f such that $g(f) = g^{SE}(\beta^{PE})$, $z^{UE}(g^{SE}(\beta^{PE})) = (\frac{\lambda}{S}, \frac{\lambda}{S})$, and*

$$\beta_p^{PE} = \left(\frac{\lambda}{S}\right)^2 \frac{4\hat{a}\tilde{a}S}{S\hat{a} + 4\tilde{a}} \tag{12}$$

Importantly, the above result yields (after some algebra) the profits of services and providers, respectively:

$$\text{Service-Profit}(s) = \tilde{a}\frac{4\tilde{a} - 3S\hat{a}}{S\hat{a} + 4\tilde{a}}\left(\frac{\lambda}{S}\right)^2 \tag{13}$$

$$\text{Provider-Profit}(p) = \lambda^2 \frac{2\hat{a}\tilde{a}}{S\hat{a} + 4\tilde{a}} \tag{14}$$

To obtain insight from these formulas, let us focus on the impact of congestion at shared versus dedicated resources. The impact of these can be seen by varying the constants $\hat{a}$ and $\tilde{a}$. If $\hat{a} \gg \tilde{a}$ ($\hat{a} \ll \tilde{a}$) then congestion is dominated by congestion at the shared (dedicated) resources.

A first observation is that the provider profit increases as $\hat{a}$ increases (i.e., shared resources become more dominant), and converges to $2S\left(\frac{\lambda}{S}\right)^2\tilde{a}$ when $\hat{a} \to \infty$. In contrast, the service profit decreases as $\hat{a}$ increases and becomes negative for large enough $\hat{a}$. Thus, *providers are profitable when shared resources dominate, but it is unprofitable for services to participate in the cloud marketplace in this setting.*

In contrast, if we let $\tilde{a}$ grow, i.e., dedicated resources dominate, then both provider and service profits increase. Further, as $\tilde{a} \to \infty$ service profits also grow unboundedly and provider profits converge to $\frac{\lambda\hat{a}}{2}$. Thus, *both services and providers are profitable if dedicated resources dominate, but services extract a dominant share of the profits*. The fact that services receive a dominant share of the profits is interesting given that competition is much larger in the service market (recall that S is the 'normalized' number of services).

Asymmetric latencies: The case of asymmetric latencies is not as simple as the symmetric case studied previously. Thus, we restrict ourselves to a particular form of asymmetry that is particularly illustrative: $\hat{a}_1 \gg \hat{a}_2, \tilde{a}_1, \tilde{a}_2$. In this case, congestion is dominated by the shared resources at provider 1, and congestion at provider 2 is in balance between shared and dedicated resources (in comparison to provider 1). Thus, there is a "dominant" provider in this case which has more efficient infrastructure. The question we look at in the analysis is the extent to which this dominant provider can exploit this increased efficiency in the marketplace.

The first step of the analysis is to characterize the provider equilibrium prices (in the same setting as Proposition 6). Using that $\hat{a}_1$ is large, we get

$$\beta_{p'}^{PE} \approx 2g_{p'} \sum_p \frac{y_p^2}{S^2 g_p^3}\tilde{a}_p, \ p' = 1, 2. \tag{15}$$

To obtain a clean characterization of the profits, we consider the case when the asymmetry becomes extreme, i.e., $\hat{a}_1 \to \infty$ and $\hat{a}_2 = \tilde{a}_1 = \tilde{a}_2 = 1$. This allows the following characterization of the profits for the services and providers.

Corollary 2. *Let $P = 2$, $\tilde{\ell}_{f_s}(x_s) = \tilde{a}x_s$ and $\hat{\ell}_p(y_p) = \hat{a}y_p$, for all s, p. Further, consider $\hat{a}_1 \to \infty$ and $\hat{a}_2 = \tilde{a}_1 = \tilde{a}_2 = 1$. Then,*

$$\begin{aligned} \textit{Provider-Profit}(1) &\to \tfrac{3}{4}S\left(\tfrac{\lambda}{S}\right)^2 \\ \textit{Provider-Profit}(2) &\to 3S\left(\tfrac{\lambda}{S}\right)^2 \\ \textit{Service-Profit}(s) &\to -\tfrac{3}{4}\left(\tfrac{\lambda}{S}\right)^2 \end{aligned} \tag{16}$$

A first comment on Corollary 2 is that it is in complete alignment with (13) and (14) for the case of symmetric providers. Specifically, in the case of symmetric providers, services are unprofitable and providers are profitable when shared resources dominate congestion. That is exactly what we see in this asymmetric setting as well.

However, Corollary 2 also highlights something interesting about the competition among providers. In particular, *the market structure protects the inefficient provider by limiting the market power of the dominant provider*. In particular, despite the fact that the gap in efficiency is extreme, provider 2 can only extract at most four times the profit of provider 1, and provider 1 still extracts significant profit. The reason for this is that, even though the latency function at provider 1 is very steep, provider 1 still manages to obtain at least 1/3 of the services (see the proof). Importantly, because of the steepness, these are services with very little user traffic; however provider 1 can still obtain significant profit from them.

Importantly, the insights provided by Corollary 2 hold more generally as well, as illustrated by Figure 2, which illustrates provider and service profits when $\hat{a}_1 < \infty$.

Though we have only considered one form of asymmetry above, this setting is quite representative. In particular, the other settings we have considered all reinforce the impact of shared vs. dedicated resources on profitability highlighted by Corollary 1 and the limited market power attainable in the provider market highlighted by Corollary 2. Thus, we omit them due to space constraints.

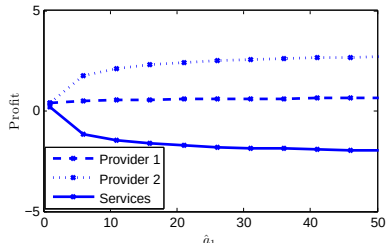

Fig. 2. Service and provider profits when providers are asymmetric.

B. Efficiency of user performance

The previous analysis focuses on the impact of the marketplace on services and providers, we now shift to the impact of the market structure on the user performance. In particular, our goal is to understand how price competition among services and providers impacts the performance experienced by users.

To perform this study, we contrast the performance at equilibrium to the performance achievable if the allocation of traffic to services and providers was performed optimally. In particular, we study the following ratio, which is often termed the "price of anarchy" [19]. To state the definition we need to first define the 'network latency', our measure for the aggregate user-experienced performance:

$$\ell(z,g) \stackrel{\text{def}}{=} \frac{1}{\lambda}\sum_p Sg_pz_p\left((\tilde{\ell}_p(z_p)+\hat{\ell}_p(Sg_pz_p)\right) \tag{17}$$

This form follows from using Lemma 1 and recalling that Sg_pz_p is the total traffic to provider p, which is normalized by the scaling parameter n in the large system limit. (Note that $\sum_p Sg_pz_p = \lambda$.)

Definition 6. *We define the* ***price of anarchy*** *(PoA) as*

$$PoA \stackrel{\text{def}}{=} \sup_{\beta^{PE}} \frac{\ell(z^{UE},g^{SE})}{\ell(z^*,g^*)} \geq 1 \tag{18}$$

where $g^{SE} = g^{SE}(\beta^{PE})$ and $z^{UE} = z^{UE}(g^{SE})$ satisfy (8) *when $\beta = \beta^{PE}$, and*

$$\begin{aligned}(z^*,g^*) \in \underset{x\geq 0, g\geq 0}{\text{argmin}} \quad & \ell(z,g) \\ \textit{s.t.:} \quad & \textstyle\sum_p g_pSz_p = \lambda \\ & \textstyle\sum_s g_p = 1.\end{aligned} \tag{19}$$

While many results on the price of anarchy of non-atomic routing games are known, the multi-tier structure of our model add significant complexity in deriving such results. To highlight the challenge, note that the mappings g^* and $g^{SE}(\beta^{SE})$ differ in general, which makes analysis difficult. However, the following straightforward lemma highlights that it is possible to bound the price of anarchy by focusing on cases with "fixed service-to-provider mappings".

Lemma 3. *$PoA \geq PoA_{g^{SE}(\beta^{PE})}$, where*

$$PoA_g \stackrel{\text{def}}{=} \frac{\ell(z^{UE}(g),g)}{\ell(z^*(g),g)} \geq 1, \forall g$$

and

$$\begin{aligned}z^*(g) \in \underset{z\geq 0}{\text{argmin}} \quad & \ell(z,g) \\ \textit{s.t.:} \quad & \textstyle\sum_p g_pSz_p = \lambda.\end{aligned} \tag{20}$$

Using this lemma, we can obtain lower bounds on the price of anarchy by studying the price of anarchy in the (simpler) case of fixed service-to-provider mappings. In this case, we can prove the following bounds.

Theorem 1. *Let $\epsilon > 0$. Consider $P = 2$, $\tilde{\ell}_p(y_p) = \tilde{a}_p y_p^k$, and $\hat{\ell}_p(y_p) = \hat{a}_p y_p^k$. Then,*

$$\Omega(k^{1-\epsilon}) \leq \sup PoA_{g^{SE}(\beta^{PE})} \leq k+1,$$

where the sup *is taken over $S, \lambda, \tilde{a}, \hat{a}, \beta^{PE}$. Thus, $PoA \geq \Omega(k^{1-\epsilon})$.*

Theorem 1 highlights that the price of anarchy grows quickly as the non-linearity of the latency functions grow. This behavior is not unexpected, as a similar result holds for classical non-atomic routing games [22]. However, the consequence in this setting is important: it highlights that *the cloud marketplace can yield equilibria that have inefficient user performance*.

Note that Theorem 1 provides a negative result; however we conjecture that positive results are also possible. In particular, we conjecture that the price of anarchy remains small in the case of linear latency functions, similarly to what is observed in classical non-atomic routing games. In fact, when $P = 2$, our numerical studies support the claim that $PoA \leq 7/6$. However, proving such a result is difficult.

V. Concluding Remarks

In this paper we develop a three-tier market model for a cloud marketplace. Our focus is on a setting including users purchasing services from SaaS providers, which in turn purchase computing resources from either PaaS or IaaS. Within each level we define and characterize competitive equilibria. Further, we use these characterizations to understand the profitability of SaaSs and PaaSs/IaaSs, and to understand the impact of price competition on user experienced performance.

The results in this paper represent a starting point for the analysis of the novel model presented here, and we hope that the model is of interest in its own right for future research. In particular, it captures a rich interaction between pricing and congestion across multiple markets; and there are many open questions that remain. For example, our results require (for technical reasons) a variety of simplifying assumptions. It would be quite interesting to relax these, e.g., study $P > 2$ and characterize equilibrium existence and uniqueness for non-linear latency functions. Additionally, our price of anarchy analysis focused on lower bounds, but upper bounds are also important (but seemingly difficult) to obtain. Finally, a particularly important future direction is to study the contrasts between different pricing schemes used by services and providers using the model in this paper.

References

[1] D. Acemoglu and A. Ozdaglar. Competition and efficiency in congested markets. *Math. Oper. Res.*, 32(1):1–31, 2007.
[2] E. Altman, U. Ayesta, and B. Prabhu. Load balancing in processor sharing systems. In *Proc. of ValueTools*, pages 1–10, 2008.
[3] E. Altman, T. Boulogne, R. El-Azouzi, T. Jiménez, and L. Wynter. A survey on networking games in telecommunications. *Comput. Oper. Res.*, 33(2):286–311, 2006.
[4] Amazon EC2 Outage Reveals Challenges of Cloud Computing). *http://cloudtimes.org/2012/07/03/amazon-outage-risk-computing/*.
[5] Amazon EC2 Pricing. *http://aws.amazon.com/ec2/pricing/*.
[6] Amazon Simple Queue Service (Amazon SQS). *http://aws.amazon.com/sqs/*.

[7] J. Anselmi, U. Ayesta, and A. Wierman. Competition yields efficiency in load balancing games. *Perf. Eval.*, 68(11):986–1001, 2011.

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

[9] D. Ardagna, B. Panicucci, and M. Passacantando. Generalized Nash Equilibria for the Service Provisioning Problem in Cloud Systems. *IEEE Trans. on Services Computing Preprint*, 2012.

[10] M. Beckmann, C. B. Mcguire, and C. B. Winsten. *Studies in the Economics of Transportation*. Yale University Press, New Haven, CT, 1956.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004.

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

[13] R. Cominetti, J. R. Correa, and N. E. Stier-Moses. The impact of oligopolistic competition in networks. *Oper. Res.*, 57:1421–1437, November 2009.

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

[15] Google App Engine Pricing. *http://cloud.google.com/pricing/*.

[16] O. D. Hart. Monopolistic competition in a large economy with differentiated commodities. *Review of Economic Studies*, 46(1):1–30, 1979.

[17] M. Haviv. The Aumann-Shapely pricing mechanism for allocating congestion costs. *Operations Research Letters*, 29(5):211–215, 2001.

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

[19] E. Koutsoupias and C. Papadimitriou. Worst-case equilibria. In *STACS*, volume 1563 of *Proc. of STOC*, pages 404–413, January 1999.

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

[21] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, New York, NY, USA, 2007.

[22] T. Roughgarden and E. Tardos. How bad is selfish routing? *J. ACM*, 49(2):236–259, Mar. 2002.

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

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

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

[26] J. G. Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers*, 1:325–378, 1952.

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

APPENDIX

A. Proof of Proposition 1

With respect to multipliers $L_{1,p}$, L_2 and $L_{3,s} \geq 0$, the KKT conditions of (3) read

$$\begin{array}{ll} \tilde{\ell}_{f_s}(x_s) + \alpha_s - L_{1,f_s} - L_2 - L_{3,s} = 0, & \forall s \\ \hat{\ell}_p(y_p) + L_{1,p} = 0, & \forall p \end{array} \tag{21}$$

plus feasibility constraints and complementarity slackness. Substituting the second equation into the former, we obtain conditions (2). Since (3) is a strictly convex optimization problem, there exists a unique minimizer; and thus a unique user equilibrium.

B. Proof of Lemma 1

For contradiction, assume that $\alpha^{SE}_{s_1} < \alpha^{SE}_{s_2}$. Then, $x^{UE}_{s_1}(\alpha^{SE}, f) > x^{UE}_{s_2}(\alpha^{SE}, f)$ by the definition of user equilibrium. Given the structure of $\alpha^{SE}_{s_1}(f)$ (5), which is proven in Proposition 5, we have that $\alpha^{SE}_{s_1}(f)$ is strictly increasing in x_{s_1}. This follows because the fraction that multiplies x_{s_1} in (5) does not change when x_{s_1} varies and $x_{s_1} + x_{s_2}$ is kept constant. Consequently, $\alpha^{SE}_{s_1} > \alpha^{SE}_{s_2}$, which is a contradiction. Further, the same argument gives that $\alpha^{SE}_{s_1} > \alpha^{SE}_{s_2}$ does not hold, and so $\alpha^{SE}_{s_1} = \alpha^{SE}_{s_2}$.

But, if $\alpha^{SE}_{s_1} = \alpha^{SE}_{s_2}$, then (5) ensures that $x_{s_1} = x_{s_2}$. Since s_1 and s_2 are general, this holds true in general, and $x_s = \frac{y_{f_s}}{Sg_{f_s}}$. Substituting $x_s = \frac{y_{f_s}}{Sg_{f_s}} = z_p$ in Definition 1, we get conditions (7). Existence and uniqueness of a vector z^{UE} that solves (7) follows easily by using the potential function method (similarly to the proof of Proposition 1).

C. Proof of Proposition 2

Consider Formula (5). If the latencies are linear, the right-hand term can be rewritten as $x_{s'}(\tilde{a}_{p'} + c_{s'})$, where $p' = f_{s'}$ and $c_{s'}$ is some constant. Let $\alpha^\star$ be such that $\alpha^\star_s$ is given by Formula (5) where the x satisfies the following conditions

$$\begin{array}{l} \min_{s':x_{s'}>0} \left\{ \tilde{\ell}_{f_{s'}}(x_{s'}) + \hat{\ell}_{f_{s'}}(y_{f_{s'}}) + x_{s'}(\tilde{a}_{f_{s'}} + c_{s'}) \right\} \\ = \tilde{\ell}_{f_s}(x_s) + \hat{\ell}_{f_s}(y_{f_s}) + x_s(\tilde{a}_{f_s} + c_s),\ \forall s : x_s > 0 \\ \sum_{s:f_s=p} x_s = y_p,\ \forall p, \\ \sum_s x_s = \lambda. \end{array} \tag{22}$$

Since $\tilde{\ell}_p(0) = \hat{\ell}_p(0) = 0$ for all p, previous conditions read

$$\begin{array}{l} \tilde{\ell}_{f_1}(x_1) + \hat{\ell}_{f_1}(y_{f_1}) + x_1(\tilde{a}_{f_1} + c_1) \\ = \tilde{\ell}_{f_s}(x_s) + \hat{\ell}_{f_s}(y_{f_s}) + x_s(\tilde{a}_{f_s} + c_s),\ \forall s > 1 \\ \sum_{s:f_s=p} x_s = y_p,\ \forall p, \\ \sum_s x_s = \lambda. \end{array} \tag{23}$$

and form a linear system with $S + P$ unknowns and $S + P$ independent equations, thus there exists a unique solution (say $x^\star$) and it is such that $x^\star_s > 0$ for all s. We claim that $\alpha^\star$, which is now well-defined because $\alpha^\star_s = x^\star_s(\tilde{a}_{f_s} + c_s)$ for all s, is a service-equilibrium price vector. To prove this, we use Definition 2 and show that no service has the incentive in changing its price. This amounts to check that for all s':

$$\begin{array}{rll} \alpha^\star_{s'} \in & \underset{\overline{\alpha}_{s'} \geq 0}{\operatorname{argmax}} & \overline{\alpha}_{s'} x^{UE}_{s'}((\overline{\alpha}_{s'}, \alpha^\star_{-s'}), f) \\ \equiv & \underset{\overline{\alpha}_{s'} \geq 0, L_{3,s}}{\operatorname{argmax}} & \overline{\alpha}_{s'} x_{s'} \\ & \text{s.t.:} & \tilde{\ell}(x_{s'}) + \hat{\ell}(y_{f_{s'}}) + \alpha^\star_{s'} - L_{3,s'} = \\ & & \tilde{\ell}(x_s) + \hat{\ell}(y_{f_s}) + \alpha^\star_s - L_{3,s}, \forall s \neq s' \\ & & \sum_{s:f_s=p} x_s = y_p,\ \forall p, \\ & & \sum_s x_s = \lambda \\ & & L_{3,s} \geq 0, \forall s \\ & & L_{3,s} x_s = 0, \forall s \end{array} \tag{24}$$

which follows by using that $x^{UE}((\overline{\alpha}_{s'}, \alpha^\star_{-s'}), f)$ is the optimizer of the strictly convex optimization problem (3) and substituting its KKT conditions in the constraints of the optimization in (24). One can check that the point $(\alpha^\star_{s'}, L_{3,s} = 0, \forall s)$ satisfies the KKT conditions of the optimization problem (24), for all s'. Note that if $L_{3,s} = 0, \forall s$, then (24) becomes a strictly concave optimization problem, which means that there exists a unique optimizer. Since (24) is non-concave (or non-convex), this allows us to say that $(\alpha^\star_{s'}, L_{3,s} = 0, \forall s)$ is a local maximum. However, we now prove that $(\alpha^\star_{s'}, L_{3,s} = 0, \forall s)$ is actually a global maximum. By contradiction, assume that $L_{3,s} > 0$ for some s. Then, $x_s = 0$ by complementarity slackness, but this implies that the profit of provider s is zero, in contrast with previous case which was positive. This proves existence. Now, the facts that $L_{3,s} = 0, \forall s$ in any stationary point of the Lagrangian of (24) and that (24) is strictly concave problem when $L_{3,s} = 0, \forall s$ prove uniqueness.

D. Proof of Proposition 3

We begin with a technical lemma.

Lemma 4. *Let f be given and $\alpha = \alpha^{SE}(f)$ be a service-equilibrium price vector. Then, $\alpha_s x_s^{UE}(\alpha, f) > 0$, $\forall s$.*

Proof: To begin, assume that $\alpha = 0$. It follows that there exists $\epsilon > 0$ such that each service s has the incentive in setting price $\alpha_s = \epsilon$. To see this, note that, by the assumption $\tilde{\ell}_{f_s}(0) = \hat{\ell}_{f_s}(0) = 0$, $\forall s$ it follows that $x_s^{UE}(0, f) > 0, \forall s$. Take $\alpha_s = \epsilon = \min_{s' \neq s} \ell_{s'}(x^{UE}(0, f)) > 0$. In this case, $\min_{s' \neq s} \ell_{s'}(x^{UE}(0, f)) < \min_{s' \neq s} \ell_{s'}(x^{UE}(\epsilon e_s, f))$, where e_s is the unit vector in direction s, because of the monotonicity of the latencies. In other words, $\epsilon < \min_{s' \neq s} \ell_{s'}(x^{UE}(\epsilon e_s, f))$, which implies $x_s^{UE}(\epsilon e_s, f) > 0$ and then $\epsilon x_s^{UE}(\epsilon e_s, f) > 0$.

Now, to finish the proof, we show that if there exists s' such that $\alpha_{s'} x_{s'}^{UE}(\alpha, f) > 0$, then $\alpha_s x_s^{UE}(\alpha, f) > 0$, $\forall s$. To prove this part, we proceed as in Lemma 4.2 of [1]. Let $K \stackrel{\text{def}}{=} \alpha_{s'} + \ell_s(x^{UE}(\alpha, f))$, which is positive by assumption. Assume $\alpha_s x_s^{UE}(\alpha, f) = 0$ for some s and consider the price $\alpha_s = K - \epsilon$ for some small $\epsilon > 0$. Using that $\tilde{\ell}_{f_s}(0) = \hat{\ell}_{f_s}(0) = 0$, necessarily $x_s^{UE}(\alpha(f), f) > 0$, which violates the hypothesis that the profit of s is zero. ∎

Using the above lemma, we can now prove Proposition 3. Let f be given and $\alpha = \alpha^{SE}(f)$. Using that $\alpha_s x_s^{UE}(\alpha, f) > 0$ for all s (by Lemma 4), service-equilibrium prices of service s' are the optimizers of the following maximization

$$\begin{array}{ll} \max\limits_{\alpha_{s'}, x \geq 0} & \alpha_{s'} x_{s'} \\ \text{s.t.:} & \tilde{\ell}_{p'}(x_{s'}) + \hat{\ell}_{p'}(y_{p'}) + \alpha_{s'} = \tilde{\ell}_{f_s}(x_s) + \hat{\ell}_{f_s}(y_{f_s}) + \alpha_s, \\ & \forall s \neq s' \\ & \sum_{s: f_s = p} x_s = y_p, \, \forall p \\ & \sum_{s=1}^{S} x_s = \lambda \end{array} \tag{25}$$

where $p' = f_{s'}$. Imposing to zero the partial derivatives of the Lagrangian of previous optimization problem, with respect to multipliers $L_{1,s}$, $L_{2,p}$, $L_3 \in \mathbb{R}$ we obtain

$$\begin{array}{ll} x_{s'} = \sum_{s \neq s'} L_{1,s} & \\ \alpha_{s'} = \sum_{s \neq s'} L_{1,s} \tilde{\ell}'_{p'}(x_{s'}) - L_{2,p'} - L_3 & \\ -L_{1,s} \tilde{\ell}_{f_s}(x_s) - L_{2,f_s} - L_3 = 0 & \forall s \neq s' \\ \sum_{s \neq s': f_s \neq p'} L_{1,s} \hat{\ell}'_{p'}(y_{p'}) + L_{2,p'} = 0 & \\ -\sum_{s \neq s': f_s = p} L_{1,s} \hat{\ell}'_{p}(y_{p}) + L_{2,p} = 0 & \forall p \neq p'. \end{array} \tag{26}$$

Substituting $L_{1,s}$ from the third equation, and expliciting $L_{2,p'}$ (respectively $L_{2,p}$) from the fourth (fifth) equation, after some algebra we get (5).

E. Proof of Proposition 4

By Lemma 1, we have

$$x_s = x_{S+s} = x_{2S+s} \cdots = x_{(n-1)S+s}, \, \forall s. \tag{27}$$

and $\alpha_s^{SE} = \alpha_{S+s}^{SE} = \alpha_{2S+s}^{SE} \cdots = \alpha_{(n-1)S+s}^{SE}$, $\forall s$. Substituting (27) in (5), we get

$$\alpha_{s'} = x_{s'} \ell'(x_{s'}) + x_{s'} \frac{c_1}{c_2}, \; \forall s' \in \{1, \ldots, S\} \tag{28}$$

where

$$c_1 \stackrel{\text{def}}{=} \frac{n \sum\limits_{s=1: f_s \neq p'}^{S} \left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}}{\left(\hat{\ell}'_{p'}(y_{p'})\right)^{-1} - n \sum\limits_{s=1: f_s \neq p'}^{S} \left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}} + 1 \tag{29}$$

$$\begin{aligned} c_2 \stackrel{\text{def}}{=} & \frac{n \sum\limits_{\substack{s=1: \\ f_s \neq p'}}^{S} \left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1} \left(n \sum\limits_{\substack{s \neq s': \\ f_s = p'}} \left(\tilde{\ell}'_{p'}(x_s)\right)^{-1} - \left(\tilde{\ell}'_{p'}(x_{s'})\right)^{-1} \right)}{\left(\hat{\ell}'_{p'}(y_{p'})\right)^{-1} - n \sum\limits_{s=1: f_s \neq p'}^{S} \left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}} \\ & + \sum_{p \neq p'} \frac{n^2 \left(\sum\limits_{s=1: f_s = p}^{S} \left(\tilde{\ell}'_{p}(x_s)\right)^{-1} \right)^2}{\left(\hat{\ell}'_{p}(y_{p})\right)^{-1} + n \sum\limits_{s=1: f_s = p}^{S} \left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1}} - \left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1} \\ & + n \sum_{s=1}^{S} \left(\tilde{\ell}'_{f_s}(x_s)\right)^{-1} \end{aligned} \tag{30}$$

and, with a slight abuse of notation, we have used $y_p = \frac{1}{n} \sum_{s=1: f_s = p}^{nS} x_s = \sum_{s=1: f_s = p}^{S} x_s$. As $n \to \infty$, $c_1 \to 0$ and $|c_2| \to \infty$, and we conclude that (6) holds.

Now, putting (6) in (2), we get the conditions

$$\begin{array}{l} \min\limits_{s': x_{s'} > 0} \left\{ \tilde{\ell}_{f_{s'}}(x_{s'}) + \hat{\ell}_{f_{s'}}(y_{f_{s'}}) + x_{s'} \tilde{\ell}'_{f_{s'}}(x_{s'}) \right\} = \\ \tilde{\ell}_{f_s}(x_s) + \hat{\ell}_{f_s}(y_{f_s}) + x_s \tilde{\ell}'_{f_s}(x_s), \forall s \in \{1, \ldots, S\} : x_s > 0 \\ \sum_{s: f_s = p} x_s = y_p, \, \forall p \\ \sum_{s=1}^{S} x_s = \lambda \\ x_s \geq 0, \, \forall s, \end{array} \tag{31}$$

which coincide with the KKT conditions of optimization problem

$$\begin{array}{ll} \min_{x \geq 0} & \sum\limits_{s=1}^{S} x_s \tilde{\ell}_{f_s}(x_s) + \sum\limits_{p=1}^{P} \int_0^{y_p} \hat{\ell}_p(z) \mathrm{d}z \\ & \sum_{s=1: f_s = p}^{S} x_s = y_p, \, \forall p \\ & \sum_{s=1}^{S} x_s = \lambda. \end{array} \tag{32}$$

Problem (32) is strictly convex because the $\tilde{\ell}_p(\cdot)$s and the $\hat{\ell}_p(\cdot)$s are increasing and convex by assumption, and therefore there exists a unique optimizer [11].

F. Proof of Proposition 5

Consider the optimization problem

$$\begin{array}{ll} \min\limits_{g \geq 0, z \geq 0} & \sum\limits_{p} z_p S g_p \tilde{\ell}_p\left(z_p\right) + \int_0^{z_p S g_p} \hat{\ell}_p(\overline{z}_p) \mathrm{d}\overline{z}_p + S \beta_p g_p \\ \text{s.t.:} & \sum_p g_p = 1, \\ & \sum_p g_p S z_p = \lambda. \end{array} \tag{33}$$

Imposing the partial derivatives of the Lagrangian of (33) to zero and with respect to multipliers $L_1, L_2, L_{3,p} \geq 0, L_{4,p} \geq 0$ for all p, after some algebra we get

$$\begin{array}{ll} S\beta_p - S z_p^2 \tilde{\ell}'_p\left(z_p\right) - L_{3,p} = L_1, & \forall p \\ \tilde{\ell}_p\left(z_p\right) + \hat{\ell}_p(z_p S g_p) + z_p \tilde{\ell}'_p\left(z_p\right) - L_{4,p} = L_2, & \forall p. \end{array} \tag{34}$$

Now, using that the $L_{3,p}$s and the $L_{4,p}$s are non-negative and that the complementarity slackness conditions ensure $L_{3,p} g_p = 0$ and $L_{4,p} y_p = 0$, conditions (34) can be written as

$$\begin{array}{l} \min\limits_{p': g_{p'} > 0} \left\{ S\beta_{p'} - S z_{p'}^2 \tilde{\ell}'_{p'}\left(z_{p'}\right) \right\} \\ = S\beta_p - S z_p^2 \tilde{\ell}'_p\left(z_p\right), \forall p : g_p > 0 \\ \\ \min\limits_{p': z_{p'} > 0} \left\{ \tilde{\ell}_{p'}\left(z_{p'}\right) + \hat{\ell}_{p'}(z_{p'} S g_{p'}) + z_{p'} \tilde{\ell}'_{p'}\left(z_{p'}\right) \right\} \\ = \tilde{\ell}_p\left(z_p\right) + \hat{\ell}_p(z_p S g_p) + z_p \tilde{\ell}'_p\left(z_p\right), \, \forall p : z_p > 0. \end{array} \tag{35}$$

Furthermore, in an optimal solution of (33), we have $z_p > 0$ for all p. In fact, assuming that $z_p = 0$ for some p, we have $L_{4,p} = -L_2 < 0$, by (34), because there always exists p' such that $z_{p'} > 0$ for which (34) and $L_{4,p} z_p = 0$ ensures that $L_2 = \tilde{\ell}_{p'}(z_{p'}) + \hat{\ell}_{p'}(z_{p'} S g_{p'}) + z_{p'} \tilde{\ell}'_{p'}(z_{p'})$, which is strictly positive because the latencies are increasing. Therefore, substituting in (35) that $z_p > 0$ for all p and since in an optimal solution $g_p > 0$ for all p (by hypothesis), we get conditions

$$\begin{aligned} &\beta_p - z_p^2 \tilde{\ell}'_p(z_p) = \beta_1 - z_1^2 \tilde{\ell}'_1(z_1), \forall p \\ &\tilde{\ell}_p(z_p) + \hat{\ell}_p(z_p S g_p) + z_p \tilde{\ell}'_p(z_p) = \\ &\tilde{\ell}_1(z_1) + \hat{\ell}_1(z_1 S g_1) + z_1 \tilde{\ell}'_1(z_1), \forall p. \end{aligned} \tag{36}$$

Conditions (36), together with the constraints of (33), coincide with the conditions (8) that define a service-equilibrium distribution if $g_p^{SE} > 0, \forall p$. Therefore, service-equilibrium distributions $g^{SE} = g^{SE}(\beta^{PE})$, provided that they exist, are exactly the optimizers of (33). With the change of variable $z_p = \frac{y_p}{S g_p}$, (33) is equivalent to (9). Using that the $\tilde{\ell}_p()$'s and the $\hat{\ell}_p()$'s are convex, one can check that the objective funtion of (9) is strictly convex (the Hessian is strictly positive semi-definite). Since (9) is defined on a non-empty and convex domain and that the g_p's are positive in an optimal solution, (9) is a strictly-convex optimization problem; see [11].

G. Proof of Lemma 2

Let β be a provider equilibrium. Then, $\beta \neq 0$. In fact, if $\beta = 0$ then $g_p^{SE}(0) = 1$ where $p = \text{argmax}_{p'} \tilde{a}_{p'}$ (this is evident from (8)). But p does not make any profit because $\beta_p = 0$ and thus has the incentive to deviate. If $\beta_{p'} > 0$ for (at least) provider p', then again the structure of (8) ensures that all the other providers make strictly positive profit (all of them have the incentive in increasing of ϵ their price), i.e., $\beta_{p'} g_{p'}^{SE}(\beta) > 0$ for all p, which implies $g_{p'}^{SE}(\beta) > 0$.

H. Proof of Proposition 6

Let price vector $\beta = (\beta_1, \beta_2)$ be a provider equilibrium. Using that in a provider equilibrium $g_p(\beta) > 0$ for all p (by Lemma 2), the values of β_p that maximize the profit of provider p are the optimizers of the following maximization problem (say PROB_p)

$$\begin{aligned} \max_{\substack{\beta_p \geq 0 \\ g \geq 0, y \geq 0}} \quad & \beta_p g_p S \\ \text{s.t.:} \quad & \tilde{a}_1 \left(\tfrac{y_1}{S g_1}\right)^2 - \beta_1 = \tilde{a}_2 \left(\tfrac{y_2}{S g_2}\right)^2 - \beta_2, \\ & g_1 + g_2 = 1, \\ & 2\tilde{a}_1 \tfrac{y_1}{S g_1} + \hat{a}_1 y_1 = 2\tilde{a}_2 \tfrac{y_2}{S g_2} + \hat{a}_2 y_2, \\ & y_1 + y_2 = \lambda \end{aligned} \tag{37}$$

Without loss of generality we study PROB_1. The Lagrangian of PROB_1, with respect to multipliers $L_i \in \mathbb{R}$, $i = 1, \ldots, 5$, reads

$$\begin{aligned} \mathcal{L} \stackrel{\text{def}}{=} \quad & -\beta_1 g_1 S + \\ & L_1 \left(\tilde{a}_1 \left(\tfrac{y_1}{S g_1}\right)^2 - \beta_1 - \tilde{a}_2 \left(\tfrac{y_2}{S g_2}\right)^2 + \beta_2 \right) + \\ & L_2 (g_1 + g_2 - 1) + \\ & L_3 \left(\hat{\ell}_1(y_1) + 2\tilde{a}_1 \tfrac{y_1}{S g_1} - \hat{\ell}_2(y_2) - 2\tilde{a}_2 \tfrac{y_2}{S g_2} \right) + \\ & L_4 (y_1 + y_2 - \lambda) \\ & -L_5 \beta_1, \end{aligned} \tag{38}$$

where $L_5 \geq 0$. Imposing the partial derivatives (with respect to $\beta_1, g_1, g_2, y_1, y_2$) of $\mathcal{L}$ to zero, we obtain that the following conditions are necessarily satisfied in a stationary point of $\mathcal{L}$:

$$\begin{aligned} & g_1 S = -L_1 - L_5 \\ & \beta_1 S = -L_1 \tfrac{2y_1^2}{S^2 g_1^3} \tilde{a}_1 + L_2 - L_3 \tfrac{y_1}{S g_1^2} 2\tilde{a}_1 \\ & L_1 \tfrac{2y_2^2}{S^2 g_2^3} \tilde{a}_2 + L_2 + L_3 \tfrac{y_2}{S g_2^2} 2\tilde{a}_2 = 0, \\ & L_1 \tfrac{2y_1}{S^2 g_1^2} \tilde{a}_1 + L_3 (\hat{a}_1 + \tfrac{2\tilde{a}_1}{S g_1}) + L_4 = 0 \\ & -L_1 \tfrac{2y_2}{S^2 g_2^2} \tilde{a}_2 - L_3 (\hat{a}_2 + \tfrac{2\tilde{a}_2}{S g_2}) + L_4 = 0. \end{aligned} \tag{39}$$

Now, assume that the optimizing β_1 is greater than zero. Then, $L_5 = 0$ by complementarity slackness conditions. Using $g_1 S = -L_1$ and substituting L_2 (respectively L_4) from the second (fifth) equation in the first (fourth), we obtain

$$\begin{aligned} & \beta_1 S = g_1 S \sum_p \tfrac{2y_p^2}{S^2 g_p^3} \tilde{a}_p - L_3 \left(\sum_p \tfrac{y_p}{S g_p^2} 2\tilde{a}_p \right) \\ & g_1 S \frac{\sum_p \tfrac{2y_p}{S^2 g_p^2} (\tilde{a}_p + c_p)}{\sum_p \hat{a}_p + \tfrac{2\tilde{a}_p + c_p}{S g_p}} = L_3 \end{aligned} \tag{40}$$

After some algebra, this yields (11). Note that this is not enough to claim that the optimizing β_1 satisfies (11) because we assumed β_1 positive. To complete the proof we need to argue that β_1 is positive.

Lemma 5. *The right-hand term of* (11) *is non-negative.*

Proof: To prove the statement, we show that

$$\begin{aligned} & \left(\tilde{a}_1 \tfrac{y_1^2}{S^2 g_1^3} + \tilde{a}_2 \tfrac{y_2^2}{S^2 g_2^3} \right) \left(\tfrac{2\tilde{a}_1}{S g_1} + \hat{a}_1 + \tfrac{2\tilde{a}_2}{S g_2} + \hat{a}_2 \right) \\ & -2S \left(\tilde{a}_1 \tfrac{y_1}{S^2 g_1^2} + \tilde{a}_2 \tfrac{y_2}{S^2 g_2^2} \right)^2 \geq 0. \end{aligned} \tag{41}$$

Considering $\hat{a}_1 = \hat{a}_2 = 0$, after some algebra the left-hand side of previous inequality becomes $(y_1/g_1 - y_2/g_2)^2$. ∎

I. Proof of Corollary 1

Using the linearity of the latency functions and substituting (11) in the definition of service-equilibrium distribution, in a provider equilibrium we have the following conditions

$$\tilde{a} \left(\frac{y_1}{S g_1}\right)^2 - \tilde{a} \left(\frac{y_2}{S g_2}\right)^2 =$$

$$2(g_1 - g_2) \left[\sum_p \tilde{a} \frac{y_p^2}{S^2 g_p^3} - \frac{2}{S} \frac{\left(\sum_p \tfrac{\tilde{a} y_p}{S g_p^2} \right)^2}{\sum_p \hat{a} + \tfrac{2\tilde{a}}{S g_p}} \right] \tag{42}$$

$$\begin{aligned} & 2\tilde{a} \frac{y_1}{S g_1} + \hat{a} y_1 = 2\tilde{a} \frac{y_2}{S g_2} + \hat{a} y_2, \\ & g_1 + g_2 = 1, \\ & y_1 + y_2 = \lambda \end{aligned} \tag{43}$$

with $g_p, y_p \geq 0$, $\forall p$. By substitution, one can check that $g_p = 1/2, y_p = \lambda/2$, $\forall p$, solves the above equations. To prove the statement, we need to show that that is the unique solution. For contradiction, assume that $y_1 > y_2$. Then, by (43), $g_1 < g_2$ and the left-hand side of equation (42) is positive. Since $\sum_p \tilde{a} \tfrac{y_p^2}{S^2 g_p^3} - \tfrac{2}{S} (\sum_p \tfrac{\tilde{a} y_p}{S g_p^2})^2 / \sum_p (\hat{a} + \tfrac{2\tilde{a}}{S g_p})$ is non-negative (by Lemma 5), the right-hand side of equation (43) becomes non-positive, i.e., a contradiction. Therefore, $g_p = 1/2, y_p = \lambda/2$, $\forall p$, is the unique solution and substituting in (11) we get (12).

J. Proof of Corollary 2

Substituting (15) in (8), after some algebra we get

$$\begin{aligned}
&\tilde{a}_1 z_1^2\left(1-\frac{2g_2}{g_1}\right)=\tilde{a}_2 z_2^2\left(1-\frac{2g_1}{g_2}\right)\\
&2\tilde{a}_1 z_1+\hat{a}_1 S g_1 z_1=2\tilde{a}_2 z_2+\hat{a}_2 S g_2 z_2 \qquad (44a)\\
&g_1+g_2=1\\
&S g_1 z_1+S g_2 z_2=\lambda, \qquad (44b)
\end{aligned}$$

which means that in a solution (y^{UE}, g^{SE}) of (44) we have

$$\frac{1}{3}\le g_p^{SE}\le\frac{2}{3},\ \forall p. \qquad (45)$$

Therefore, even though the latency at provider one is very large, competition does not prevent it in hosting at least 1/3 of the services, though the resulting traffic $S g_1 z_1^{UE}$ will be small because (44a) can be rewritten as (using (44b))

$$S g_1 z_1^{UE}=\frac{\frac{2\tilde{a}_2}{S g_2}+\hat{a}_2}{\frac{2\tilde{a}_1}{S g_1}+\hat{a}_1+\frac{2\tilde{a}_2}{S g_2}+\hat{a}_2}\,\lambda, \qquad (46)$$

which approaches zero when $\hat{a}_1\to\infty$.

Assume that $\hat{a}_1=a\to\infty$ and $\hat{a}_2=\tilde{a}_1=\tilde{a}_2=1$. The solutions of the non-linear system that is obtained by substituting (11) and (6) in (8) can be expressed in terms of a polynomial of order four. In particular, using Maple, one obtains that in a solution of such non-linear system, $g_1^{SE}=z^*$ where z^* is a root, in $[0,1]$, of the fourth-order polynomial

$$\begin{aligned}
&(3a^2\tfrac{S^2}{\lambda^2}+O(a))z^4-(a^2\tfrac{S^2}{\lambda^2}+O(a))z^3\\
&-(6a\tfrac{S}{\lambda}+O(1))z^2+(36\tfrac{S}{\lambda}+9\tfrac{S^2}{\lambda^2}+24+O(a^{-1}))z\\
&-8-8\tfrac{S}{\lambda}-2\tfrac{S^2}{\lambda^2}+O(a^{-1})=0.
\end{aligned}$$

Dividing by a^2, one gets that the solutions of this polynomial converge to the solutions of $(3z-1)z^3=0$ as $a\to\infty$. Here, all four solutions are apparently feasible, but the one of interest is $z^*=1/3$ because of (45). Using $g_1^{SE}=1/3$ and that $z_1^{UE}\to 0$ (see (46)) we get $\beta_p^{PE}=g_p^{SE}\frac{27}{4}\frac{\lambda^2}{S^2}$ and substituting in the definitions of service and provider profits, we get the expressions in the statement.

K. Proof of Theorem 1

Let $g=g^{SE}$ be a service equilibrium distribution (defined with respect to generic provider prices). The following inequalities proves $PoA_g\le k+1$ with respect to any model instance:

$$\begin{aligned}
&\ell(z^{UE}(g),g)\\
&=\sum_{p:g_p>0} S g_p z_p^{UE}\left(\tilde{a}_p(z_p^{UE})^k+\hat{a}_p(S g_p z_p^{UE})^k\right)\\
&\le(k+1)\sum_{p:g_p>0} S g_p z_p^{UE}\tilde{a}_p(z_p^{UE})^k+\int_0^{S g_p z_p^{UE}}\hat{a}_p z^k\mathrm{d}z\\
&\le(k+1)\sum_{p:g_p>0} S g_p z_p^{*}\tilde{a}_p(z_p^{*})^k+\int_0^{S g_p z_p^{*}}\hat{a}_p z^k\mathrm{d}z \qquad (47a)\\
&\le(k+1)\sum_{p:g_p>0} S g_p\tilde{a}_p(z_p^{*})^{k+1}+\hat{a}_p S g_p(z_p^{*})^{k+1}\\
&=(k+1)\,\ell(z^*,g).
\end{aligned}$$

In (47a), we use that $z^{UE}(g)$ minimizes $\sum_{p:g_p>0} S g_p z_p\tilde{\ell}_p(z_p)+\int_0^{S g_p z_p}\hat{\ell}_p(z)\mathrm{d}z$. This follows because the conditions (7) that define $z^{UE}(g)$ when $\alpha_p^{SE}=z_p\tilde{\ell}_p'(z_p)$ (recall that g is a service-equilibrium distribution), i.e., the service-equilibrium prices in the large-system limit, are the KKT conditions of the strictly-convex optimization problem

$$\begin{aligned}
\min_{z\ge 0}\quad & \textstyle\sum_{p:g_p>0} S g_p z_p\tilde{\ell}_p(z_p)+\int_0^{S g_p z_p}\hat{\ell}_p(z)\mathrm{d}z\\
\text{s.t.:}\quad & \textstyle\sum_{p:g_p>0} S g_p z_p=\lambda.
\end{aligned} \qquad (48)$$

Moving to the lower bound, we prove that $\Omega(k^{1-\epsilon})\le\sup_{S,\lambda,\tilde{a}_p,\hat{a}_p,p=1,2} PoA_g$. To do this, we consider the exhibit a bad example. Our example is the following: $\tilde{a}_1=\hat{a}_2=\delta>0$, $\hat{a}_1=2^k\,\tilde{a}_2=1$.

Using the KKT conditions of (48), which are the defining conditions of $z^{UE}(g)$, and (20), we get

$$\begin{aligned}
S g_1 z_1^{UE}&=\lambda\frac{\left(\frac{(1+k)\tilde{a}_2}{S^k g_2^k}+\hat{a}_2\right)^{1/k}}{\left(\frac{(1+k)\tilde{a}_1}{S^k g_1^k}+\hat{a}_1\right)^{1/k}+\left(\frac{(1+k)\tilde{a}_2}{S^k g_2^k}+\hat{a}_2\right)^{1/k}}\\
&\xrightarrow[\delta\to 0]{}\lambda\frac{\frac{(1+k)^{1/k}}{2g_2 S}}{1+\frac{(1+k)^{1/k}}{2g_2 S}}
\end{aligned}$$

$$S g_1 y_1^*=\lambda\frac{\left(\frac{\tilde{a}_2}{S^k g_2^k}+\hat{a}_2\right)^{1/k}}{\left(\frac{\tilde{a}_1}{S^k g_1^k}+\hat{a}_1\right)^{1/k}+\left(\frac{\tilde{a}_2}{S^k g_2^k}+\hat{a}_2\right)^{1/k}}\xrightarrow[\delta\to 0]{}\lambda\frac{\frac{1}{2g_2 S}}{1+\frac{1}{2g_2 S}}.$$

Finally, substituting into the definition of PoA_g we obtain

$$\begin{aligned}
PoA_g(k)&\xrightarrow[\delta\to 0]{}\frac{\left(\frac{\frac{(1+k)^{1/k}}{2Sg_2}}{1+\frac{(1+k)^{1/k}}{2Sg_2}}\right)^{k+1}+\frac{1}{2^k S^k g_2^k}\left(\frac{\frac{1}{1+\frac{(1+k)^{1/k}}{2Sg_2}}}{}\right)^{k+1}}{\left(\frac{\frac{1}{2Sg_2}}{1+\frac{1}{2Sg_2}}\right)^{k+1}+\frac{1}{2^k S^k g_2^k}\left(\frac{1}{1+\frac{1}{2Sg_2}}\right)^{k+1}}\\
&=\frac{2Sg_2+(1+k)^{1+1/k}}{2Sg_2+1}\left(\frac{2Sg_2+1}{2Sg_2+(1+k)^{1/k}}\right)^{k+1}\\
&\ge\frac{2Sg_2+k}{2Sg_2+1}\left(\frac{2Sg_2+1}{2Sg_2+(1+k)^{1/k}}\right)^{k+1},\\
&\approx\frac{k^{\frac{2Sg_2}{2Sg_2+1}}}{2Sg_2+1},
\end{aligned} \qquad (49)$$

where the last inequality uses that $(1+k)^{1+1/k}\ge k$. Now, by symmetry, one may have chosen the case $\tilde{a}_2=\hat{a}_1=\delta>0$, $\hat{a}_2=2^k\,\tilde{a}_1=1$ and obtained again (49) written with respect to g_1 instead of g_2. Therefore,

$$\lim_{k\to\infty}\sup_{\tilde{a}_p,\hat{a}_p,p=1,2} PoA_g(k)\times\left(\max_p\frac{k^{\frac{2Sg_p}{2Sg_p+1}}}{2Sg_p+1}\right)^{-1}\ge 1. \qquad (50)$$

Now, let $\epsilon>0$. Since there exists S such that $\max_p\frac{2Sg_p}{2Sg_p+1}>1-\epsilon$, the proof is finished.