Theoretical Computer Science 496 (2013) 113–124

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Economic models for cloud service markets: *Pricing and Capacity planning*

Ranjan Pal *, Pan Hui

University of Southern California, USA
Deutsch Telekom Laboratories, Germany

ARTICLE INFO

Keywords:
Cloud markets
Competition
Nash equilibrium
Capacity
Single-tier
Multi-tier

ABSTRACT

Cloud computing is a paradigm that has the potential to transform and revolutionalize the next generation IT industry by making software available to end-users as a service. A cloud, also commonly known as a cloud network, typically comprises of hardware (network of servers) and a collection of softwares that is made available to end-users in a *pay-as-you-go* manner. Multiple public cloud providers (e.g., Amazon) co-existing in a cloud computing market provide similar services (software as a service) to its clients, both in terms of the nature of an application, as well as in quality of service (QoS) provision. The decision of whether a cloud hosts (or finds it profitable to host) a service in the long-term would depend jointly on the price it sets, the QoS guarantees it provides to its customers, and the satisfaction of the advertised guarantees. In the *first part* of the paper, we devise and analyze three *inter-organizational* economic models relevant to cloud networks. We formulate our problems as *non cooperative* price and QoS games between *multiple* cloud providers existing in a cloud market. We prove that a *unique* pure strategy Nash equilibrium (NE) exists in two of the three models. Our analysis paves the path for each cloud provider to know what prices and QoS level to set for end-users of a given service type, such that the provider could exist in the cloud market.

A cloud provider services end-user requests on behalf of cloud customers, and due to the uncertainty in user demands over time, tend to over-provision resources like CPU, power, memory, storage, etc., in order to satisfy QoS guarantees. As a result of over-provisioning over long timescales, server utilization is very low and the cloud providers have to bear unnecessarily wasteful costs. In this regard, the price and QoS levels set by the CPs drive the end-user demand, which plays a major role in CPs estimating the *minimal capacity* to meet their advertised guarantees. By the term 'capacity', we imply the ability of a cloud to process user requests, i.e., number of user requests processed per unit of time, which in turn determine the amount of resources to be provisioned to achieve a required capacity. In the *second part* of this paper, we address the capacity planning/optimal resource provisioning problem in *single-tiered* and *multi-tiered* cloud networks using a techno-economic approach. We develop, analyze, and compare models that cloud providers can adopt to provision resources in a manner such that there is *minimum* amount of resources wasted, and at the same time the user service-level/QoS guarantees are satisfied.

Published by Elsevier B.V.

1. Introduction

Cloud computing is a type of Internet-based computing, where shared resources, hardware, software, and information are provided to end-users in an *on demand* fashion. It is a paradigm that has the potential to transform and revolutionalize

* Corresponding author at: University of Southern California, USA.
E-mail addresses: rpal@usc.edu (R. Pal), pan.hui@telekom.de (P. Hui).

0304-3975/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.tcs.2012.11.001

the IT industry by making software available to end-users as a service [1]. A public cloud typically comprises of hardware (network of servers) and a collection of softwares that is made available to the general public in a *pay-as-you-go* manner. Typical examples of companies providing public clouds include *Amazon, Google, Microsoft, E-Bay,* and commercial banks. Public cloud providers usually provide Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). The advantage of making software available as a service is three-fold [1], (1) the service providers benefit from simplified software installation, maintenance, and centralized versioning, (2) end-users can access the software in an 'anytime anywhere' manner, can store data safely in the cloud infrastructure, and do not have to think about provisioning any hardware resource due to the illusion of infinite computing resources available on demand, and (3) end-users can pay for using computing resources on a short-term basis (e.g., by the hour or by the day) and can release the resources on task completion. Similar benefit types are also obtained by making both, platform as well as infrastructure available as service.

Cloud economics will play a vital role in shaping the cloud computing industry of the future. In a recent Microsoft white paper titled "Economics of the Cloud", it has been stated that the computing industry is moving towards the cloud driven by three important economies of scale: (1) large data centers can deploy computational resources at significantly lower costs than smaller ones, (2) demand pooling improves utilization of resources, and (3) multi-tenancy lowers application maintenance labor costs for large public clouds. The cloud also provides an opportunity to IT professionals to focus more on technological innovation rather than thinking of the budget of "keeping the lights on". The economics of the cloud can be thought of having two dimensions: (1) intra-organization economics and (2) inter-organization economics. Intra-organization economics deals with the economics of internal factors of an organization like labor, power, hardware, security, etc., whereas inter-organization economics refers to the economics of market competition factors between organizations. Examples of some popular factors are price, QoS, reputation, and customer service. In this paper, we focus on inter-organizational economic issues.

Multiple public cloud providers (e.g., *Amazon, Google, Microsoft,* etc.,) co-existing in a cloud computing market provide similar services (software as a service, e.g., *Google Docs* and *Microsoft Office Live*) to its clients, both in terms of the nature of an application, as well as in quality of service (QoS) provision. The decision of whether a cloud hosts (or finds it profitable to host) a service in the long-term would (amongst other factors) depend jointly on the price it sets, the QoS guarantees it provides to its customers,[1] and the satisfaction of the advertised guarantees. Setting high prices might result in a drop in demand for a particular service, whereas setting low prices might attract customers at the expense of lowering cloud provider profits. Similarly, advertising and satisfying high QoS levels would favor a cloud provider (CP) in attracting more customers. The price and QoS levels set by the CPs thus drive the end-user demand, which, apart from determining the market power of a CP also plays a major role in CPs estimating the minimal resource capacity to meet their advertised guarantees. By the term 'capacity', we imply the ability of a cloud to process user requests, i.e., number of user requests processed per unit of time. The estimation problem is an important challenge in cloud computing with respect to resource provisioning because a successful estimation would prevent CPs to provision for the peak, thereby reducing resource wastage.

The competition in prices and QoS amongst the cloud providers entails the formation of non-cooperative games amongst competitive CPs. Thus, we have a *distributed system* of CPs (players in the game), where each CP wants to maximize its own profits and would tend towards playing a Nash equilibrium[2] (NE) strategy (i.e., each CP would want to set the NE prices and QoS levels), whereby the whole system of CPs would have no incentive to deviate from the Nash equilibrium point, i.e., the vector of NE strategies of each CP. However, for each CP to play a NE strategy, the latter should mathematically exist. In the *first part* of the paper, we address the important problem of Nash Equilibrium characterization of *different types* of price and QoS games relevant to cloud networks, its properties, practical implementability (convergence issues), and the sensitivity analysis of NE price/QoS variations by any CP on the price and QoS levels of other CPs. Our problem is important from a resource provisioning perspective as mentioned in the previous paragraph, apart from it having obvious strategic importance on CPs in terms of sustenance in the cloud market. In the second part of our paper we develop and analyze models that will be useful to cloud providers to provision resources in a manner such that there is minimum amount of resources wasted, and at the same time the user service-level/QoS guarantees are satisfied.

1.1. Related work

In regard to market competition driven network pricing, there exists research work in the domain of multiple ISP interaction and tiered Internet services [2,3], as well as in the area of resource allocation and Internet congestion management [4–6]. However, the market competition in our work relates to optimal capacity planning and resource provisioning in clouds. There is the seminal work by Songhurst and Kelly [7] on pricing schemes based on QoS requirements of users. Their work address multi-service scenarios and derive pricing schemes for each service based on the QoS

[1] A cloud provider generally gets requests from a cloud customer, which in turn accepts requests from Internet end-users. Thus, typically, the clients/customers of a cloud provider are the cloud customers. However, for modeling purposes, end-users could also be treated as customers. (See Section 2.)

[2] A group of players is in Nash equilibrium if each one is making the best decision (strategy) that he or she can, taking into account the decisions of the others.

requirements for each, and in turn bandwidth reservations. This work resembles ours to some extent in the sense that the price and QoS determined can determine optimal bandwidth provisions. However, it does not account for market competition between multiple providers and only focus on a single service provider providing multiple services, i.e., the paper addresses an intra-organization economics problem. However, in this paper, we assume single-service scenarios by multiple service providers. In a recent work [8], the authors propose a queueing driven game-theoretic model for price–QoS competition amongst multiple service providers. The work analyzes a duopolistic market between two service providers, where providers first fix their QoS guarantees and then compete for prices. Our work extends the latter cited work in the following aspects: (1) we generalize our model to incorporate n service providers, (2) we address two additional game models which are of practical importance, i.e., price–QoS simultaneous competition and prices fixed first, followed by QoS guarantees competition, (3) we provide an efficient technique to compute multiple equilibria in games, and (4) our models explicitly characterize percentile performance of parameters, which is *specific* to cloud networks provisioning resources on a percentile basis. We also want to emphasize the fact that research on price/QoS competition amongst organizations is not new in the economics domain. However, in this paper we model networking elements in price/QoS games via a queueing theoretic approach and analyze certain price/QoS games that are mainly characteristic of Internet service markets.

Recent research efforts on cloud resource provisioning have devised static and dynamic provisioning schemes. Static provisioning [19,20] is usually conducted offline and occurs on monthly or seasonal timescales,[3] whereas dynamic provisioning [21,22] dynamically adjusts to workload fluctuations over time. In both the static and the dynamic case, virtual machine (VM) sizing [1] is identified as the most important step, where VM sizing refers to the estimation of the amount of resources to be allocated to a VM or jointly to many VMs [23]. However, none of the above cited works have accounted for external factors such as cloud provider price competition, in determining the optimal capacity of a cloud provider for a given time-slot. Market competition between cloud providers is a vital factor in capacity planning because cloud providers set prices to primarily to make profits and the prices they set influence demands from end-users, and user demands drive the provisioning of optimal capacities. Other factors like scheduling policies (e.g., FCFS, Processor Sharing, etc.) employed by cloud providers, as well as the number of tiers a web application needs for service, also contribute to optimal capacity provisioning. Recent works on cloud network provisioning have accounted for parameters like scheduling and multi-tier services [24], but do not provide any analytical results on the impact of these parameters on optimally provisioned capacity, nor do they evaluate the optimal provisioned capacity. In contrast with existing approaches, we take a techno-economic approach to evaluating the optimal provisioned capacity and provide theoretical insights for our problem. Our optimal provisioned capacity is metricized by the number of user requests processed per unit of time. However, this notion of capacity can be mapped to physical resource capacity metrics like bandwidth, CPU, etc. Our proposed models aim to focus on how certain technical and economic parameters influence optimal provisioned capacity of a cloud provider, as well as other competing cloud providers, which is important when it comes to network design.

1.2. Contributions statement

Our proposed theory analyzes a few basic inter-organizational economic models through which cloud services *could* be priced under market competition. The evolution of commercial public cloud service markets is still in its inception. However, with the gaining popularity of cloud services, we expect a big surge in public cloud services competition in the years to come. The models proposed in this paper take a substantial step in highlighting relevant models to the cloud networking community for them adopt so as to appropriately price current and future cloud services. In practice, scenarios of price and/or QoS competition between organizations exist in the mobile network services and ISP markets. For example, AT&T and Verizon are competing on service, i.e., Verizon promises to provide better coverage to mobile users than AT&T, thereby increasing its propensity to attract more customers. Similarly, price competition between ISPs always existed for providing broadband services at a certain given bandwidth guarantee. Regarding our work, we also want to emphasize (1) we do not make any claims about our models being the only way to model inter-organizational cloud economics[4] and (2) there is a dependency between intra-organizational and inter-organizational economic factors, which we do not account in this paper due to modeling simplicity. However, through our work, we definitely provide readers with a concrete modeling intuition to go about addressing problems in cloud economics. *To the best of our knowledge, we are the first to provide an analytical model on inter-organizational cloud economics.*

Our Contributions — We make the following contributions in this paper.

1. We formulate a *separable end-user demand function* for each cloud provider w.r.t. to price and QoS levels set by them and derive their individual utility functions (profit function). We then define the various price–QoS games that we analyze in the paper. (See Section 2.)
2. We develop a model where the QoS guarantees provided by public CPs to end-users for a particular application type are pre-specified and fixed, and the cloud providers compete for prices. We formulate a non-cooperative price game amongst

[3] Several cloud management softwares like VMWare Capacity Planner, CapacityIQ, and IBM WebSphere CloudBurst adopt this functionality.

[4] We only model price and QoS as parameters. One could choose other parameters (in addition to price and QoS, which are essential parameters) and a different analysis mechanism than ours to arrive at a different model.

the players (i.e., the cloud providers) and prove that there exists a unique Nash equilibrium of the game, and that the NE could be practically computed (i.e., it converges). (See Section 3.)

3. We develop a non-cooperative game-theoretic model where public cloud providers *jointly* compete for the price and QoS levels related to a particular application type. We show the existence and convergence of Nash equilibria. (See Section 4.) As a special case of this model, we also analyze the case where prices charged to Internet end-users are pre-specified and fixed, and the cloud providers compete for QoS guarantees only. The models mentioned in contributions 3 and 4 drive optimal capacity planning and resource provisioning in clouds, apart from maximizing CP profits. (See Section 4.)
4. We conduct a sensitivity analysis on various parameters of our proposed models, and study the effect of changes in the parameters on the equilibrium price and QoS levels of the CPs existing in a cloud market. Through a sensitivity analysis, we infer the effect of price and QoS changes of cloud providers on their respective profits, as well as the profits of competing CPs. (See Sections 3 and 4.)[5]
5. We develop an optimization framework for single-tiered and multi-tiered cloud networks to compute the optimal provisioned capacity once the equilibrium price and QoS levels for each CP have been determined. (See Section 5.)

2. Problem setup

We consider a market of n competing cloud providers, where each provider services application types to end-users at a given QoS guarantee. We assume that end-users are customers of cloud providers in an *indirect* manner, i.e., Internet end-users use online softwares developed by companies (cloud customers), that depend on cloud providers to service their customer requests. Each CP is in competition with others in the market for services provided on the *same type* of application w.r.t functionality and QoS guarantees. For example, Microsoft and Google might both serve a word processing application to end-users by providing similar QoS guarantees. Here, the word processing application represents a particular 'type'. For a given application type, we assume that each end user signs a contract with a particular CP for a given time period,[6] and within that period it does not switch to any other CP for getting service on the same application type. Regarding contracts between a CP and its end-users, we assume that a cloud customer forwards service requests to a cloud provider on behalf of end-users, who sign up with a cloud customer (CC) for service. The CP charges its cloud customer, who in turn charges its end-users. We approximate this two-step charging scheme by modeling a virtual one-step scheme, where a CP charges end-users directly.[7]

In a given time period, each CP i positions itself in the market by selecting a price p_i and a QoS level s_i related to a given application type. Throughout the paper, we assume that the CPs compete on a single given type.[8] We define s_i as the difference between a benchmark response time upper bound, $\overline{rt}$, and the actual response time rt_i, i.e., $s_i = \overline{rt} - rt_i$. For example, if for a particular application type, every CP would respond to an end-user request within 10 s, $\overline{rt} = 10$. The response time rt_i may be defined, either in terms of the expected steady state response time, i.e., $rt_i = E(RT_i)$, or in terms of ϕ-percentile performance, $rt_i(\phi)$, where $0 < \phi < 1$. Thus, in terms of ϕ-percentile performance,[9] $P(RT_i < rt_i(\phi)) = \phi$.

We model each CP i as an M/M/1 queueing system, where end-user requests arrive as a Poisson process with mean rate λ_i, and gets serviced at a rate μ_i. We adopt an M/M/1 queueing system because of three reasons: (1) queueing theory has been traditionally used in request arrival and service problems, (2) for our problem, assuming an M/M/1 queueing system ensures tractable analyses procedures that entails deriving nice closed form expressions and helps understand system insights in a non-complex manner, without sacrificing a great deal in capturing the real dynamics of the actual arrival–departure process, and (3) the Markovian nature of the service process helps us generalize expected steady state analysis and percentile analysis together. According to the theory of M/M/1 queues, we have the following standard results [17].

$$rt_i = \frac{1}{\mu_i - \lambda_i}, \tag{1}$$

$$rt_i(\phi) = \frac{ln(\frac{1}{1-\phi})}{\mu_i(\phi) - \lambda_i}, \tag{2}$$

$$\mu_i = \lambda_i + \frac{1}{rt_i}, \tag{3}$$

[5] We study Nash equilibrium convergence as it proves the achievability of an equilibrium point in the market. We emphasize here that the existence of Nash equilibrium does not imply achievability as it may take the cloud market an eternity to reach equilibrium, even though there may exist one theoretically.

[6] In this paper, the term 'time-period' refers to the time duration of a contract between the CP and end-users.

[7] We assume here that prices are negotiated between the CP, CC, and end-users and there is a virtual direct price charging connection between the CP and its end-users. We make this approximation for modeling simplicity.

[8] In reality, each CP may in general service several application types *concurrently*. We do not model this case in our paper and leave it for future work. The case for single application types gives interesting results, which would prove to be useful in analyzing the multiple concurrent application type scenario.

[9] As an example, in cloud networks we often associate provisioning power according to the 95th percentile use. Likewise, we could also provision service capacity by accounting for percentile response time guarantees.

and

$$\mu_i(\phi) = \lambda_i + \frac{ln(\frac{1}{1-\phi})}{rt_i(\phi)}. \tag{4}$$

Eqs. (2) and (4) follow from the fact that for M/M/1 queues, $P(RT_i < rt_i(\phi)) = \phi = 1 - e^{-(\mu_i - \lambda_i rt_i(\phi))}$. Without loss of generality, in subsequent sections of this paper, we conduct our analysis on expected steady state parameters. As mentioned previously, due to the Markovian nature of the service process, the case for percentiles is exactly similar to the case for expected steady state analysis, the only difference in analysis being due to the constant, $ln(\frac{1}{1-\phi})$. Thus, all our proposed equilibrium related results hold true for percentile analysis as well.

Each cloud provider i incurs a fixed cost c_i per user request served and a fixed cost ρ_i per unit of service capacity provisioned. c_i arises due to the factor λ_i in Eq. (3) and ρ_i arises due to the factor $\frac{1}{rt_i}$ in the same equation. In this sense, our QoS-dependent pricing models are *queueing-driven*. A cloud provider charges pr_i to service each end-user request, where $pr_i \in [pr_i^{min}, pr_i^{max}]$. It is evident that each CP selects a price that results in it accruing a non-negative gross profit margin. The gross profit margin for CP i is given as $pr_i - c_i - \rho_i$, where $c_i + \rho_i$ is the marginal cost per unit of end-user demand. Thus, the price lower bound, pr_i^{min}, for each CP i is determined by the following equation.

$$pr_i^{min} = c_i + \rho_i, \quad \forall i = 1, \ldots, n. \tag{5}$$

We define the demand of any CP i, λ_i, as a function of the vectors $\mathbf{pr} = (pr_1, \ldots, pr_n)$ and $\mathbf{s} = (s_1, \ldots, s_n)$. Mathematically, we express the demand function as

$$\lambda_i = \lambda_i(\mathbf{pr}, \mathbf{s}) = x_i(s_i) - y_i pr_i - \sum_{j \neq i} \alpha_{ij}(s_j) + \sum_{j \neq i} \beta_{ij} pr_j, \tag{6}$$

where $x_i(s_i)$ is an increasing, concave, and thrice differentiable function in s_i satisfying the property of non-increasing marginal returns to scale, i.e., equal-sized reductions in response time results in progressively smaller increases in end-user demand. The functions α_{ij} are assumed to be non-decreasing and differentiable. A typical example of a function fitting $x_i(s_i)$ and $\alpha_{ij}(s_j)$ is a logarithmic function. We model Eq. (6) as a separable function of price and QoS vectors, for ensuring tractable analyses as well as for extracting independent effects of price and QoS changes on the overall end-user demand. Intuitively, Eq. (6) states that QoS improvements by a CP i result in an increase in its end-user demand, whereas QoS improvements by other competitor CPs result in a decrease in its demand. Similarly, a price increase by a CP i results in a decrease in its end-user demand, whereas price increases by other competing CPs result in an increase in its demand. Without loss of practical generality, we also assume (1) a uniform increase in prices by all n CPs cannot result in an increase in any CP's demand volume, and (2) a price increase by a given CP cannot result in an increase in the market's aggregate end-user demand. Mathematically, we represent these two facts by the following two relationships.

$$y_i > \sum_{j \neq i} \beta_{ij}, \quad i = 1, \ldots, n \tag{7}$$

and

$$y_i > \sum_{j \neq i} \beta_{ji}, \quad i = 1, \ldots, n. \tag{8}$$

The long run average profit for CP i in a given time period, assuming that response times are expressed in terms of expected values, is a function of the price and QoS levels of CPs, and is given as

$$P_i(\mathbf{pr}, \mathbf{s}) = \lambda_i(pr_i - c_i - \rho_i) - \frac{\rho_i}{\overline{rt} - s_i}, \quad \forall i. \tag{9}$$

The profit function for each CP acts as its *utility/payoff function* when it is involved in price and QoS games with other competing CPs. We assume in this paper that the profit function for each CP is known to other CPs, but none of the CPs know the values of the parameters that other competing CPs adopt as their strategy.

Problem Statement: Given the profit function for each CP (public information), how would each advertise its price and QoS values (without negotiating with other CPs) to end-users so as to maximize its own profit. In other words, in a competitive game of profits played by CPs, is there a situation where each CP is happy with its (price, QoS) advertised pair and does not benefit by a positive or negative deviation in the values of the advertised pair.

In this paper, we study games involving price and QoS as the primary parameters, i.e., we characterize and analyze the *existence*, *uniqueness*, and *convergence* of Nash equilibria. Our primary goal is to compute the optimal price and QoS levels offered by CPs to its end-users under market competition. Our analysis paves the path for each cloud provider to (1) know what price and QoS levels to set for its clients (end-users) for a given application type, such that it could exist in the cloud market, and (2) practically and dynamically provision appropriate capacity for satisfying advertised QoS guarantees, by taking advantage of the property of *virtualization* in cloud networks. The property of virtualization entails each CP to allocate optimal resources dynamically in a fast manner to service end-user requests. Using our pricing framework, in each time period, cloud providers set the appropriate price and QoS levels after competing in a game; the resulting prices drive end-user demand; the CPs then allocate optimal resources to service demand.

Table 1
List of symbols and their meaning.

Symbol	Meaning
$U_i = P_i$	Utility function of CP i
pr_i	Price charged by CP i per end-user
$\mathbf{pr}$	Price vector of CPs
pr^*	Nash equilibrium price vector
c_i	Cost incurred by CP i to service each user
λ_i	Arrival rate of end-users to CP i
ρ_i	Cost/unit of capacity provisioning by CP i
$\overline{rt}$	Response time upper bound guarantee
rt_i	Response time guarantee by CP i
C_i	Capacity cost of CP i for provisioning its user demands
ϕ	Percentile parameter
s_i	QoS level guarantee provided by CP i to its users
$\mathbf{s}$	QoS vector of CPs
s^*	Nash equilibrium QoS vector
$x_i()$	Increasing, concave, and a thrice differentiable function
$\alpha_{ij}()$	Non-decreasing and differentiable function

Remark. We decided to not analyze a competitive market, i.e., where CPs are price/QoS taking and a Walrasian equilibrium results when demand equals supply, because a competitive market analysis is mainly applicable when the resources traded by an organization are negligible with respect to the total resource in the system [9,10]. In a cloud market this is definitely not the case as there are a few cloud providers and so the resource traded by one is *not* negligible with respect to the total resources traded in the system. Therefore we analyze oligopolistic markets where CPs are price/QoS anticipating.

We consider the following types of price–QoS game models in our work.

1. CP QoS guarantees are pre-specified; CPs compete with each other for prices, given QoS guarantees. (Game 1)
2. CPs compete for price and QoS simultaneously. (Game 2)
3. CP price levels are pre-specified; CPs compete for QoS levels. (Game 3). Game 3 is a special case of Game 2 and in Section 4, we will show that it is a Game 2 derivative.

List of Notations: For reader simplicity, we provide a table of most used notations related to the analysis of games in this paper (see Table 1).

3. Game 1—price game

In this section, we analyze the game in which the QoS guarantees of CPs are exogenously specified and the CPs compete for prices.

Game description

Players: Individual cloud providers; *Game type:* non-cooperative, i.e., no interaction between CPs; *Strategy space:* choosing a price in range $[pr_i^{min}, pr_i^{max}]$; *Player goal:* to maximize its individual utility $U_i = P_i$.

Our first goal is to show that this game has a unique price Nash equilibrium, pr^*(an instance of vector $\mathbf{pr}$), which satisfies the following first order condition

$$\frac{\partial P_i}{\partial pr_i} = -y_i(pr_i - c_i - \rho_i) + \lambda_i, \quad \forall i, \tag{10}$$

which in matrix notation can be represented as

$$\mathbf{M} \cdot \mathbf{pr} = \overline{\mathbf{x}}(\mathbf{s}) + \mathbf{z}, \tag{11}$$

where $\mathbf{M}$ is an $n \times n$ matrix with $M_{ii} = 2y_i$, $M_{ij} = -\beta_{ij}$, $i \neq j$, and where $z_i = y_i(c_i + \rho_i)$.

We have the following theorem and corollary regarding equilibrium results for our game. The readers are referred to the Appendix for the proofs.

Theorem 1. *Given that the QoS guarantees of CPs are exogenously specified, the price competition game has a unique Nash equilibrium, pr^*, which satisfies Eq. (11). The Nash equilibrium user demand, λ_i^*, for each CP i evaluates to $y_i(pr_i^* - c_i - \rho_i)$, and the Nash equilibrium profits, P_i^*, for each CP i is given by $y_i(pr_i^* - c_i - \rho_i)^2 - \frac{\rho_i}{\overline{rt} - s_i}$.*

Corollary 1. (a) *pr^* and λ^* are increasing and decreasing respectively in each of the parameters $\{c_i, \rho_i, \ i = 1, 2, \ldots, n\}$, and* (b) $\frac{\partial pr_i^*}{\partial s_j} = \frac{1}{y_i}\frac{\partial \lambda_i^*}{\partial s_j} = (M^{-1})_{ij}x_j'(s_j) - \sum_{l \neq j}(M^{-1})_{il}x_{lj}'(s_j)$.

Corollary 1 implies that (1) under a larger value for CP i's degree of positive externality δ_i, it is willing to make a bolder price adjustment to an increase in any of its cost parameters, thereby maintaining a larger portion of its original profit margin. The reason is that competing CPs respond with larger price themselves, and (2) there exists a critical value $0 \leq s_{ij}^0 \leq \overline{rt}$ such that as CP j increases its QoS level, pr_i^* and λ_i^* are increasing on the interval $[0, s_{ij}^0)$, and decreasing in the interval $[s_{ij}^0, \overline{rt})$.

Sensitivity analysis: We know the following relationship

$$\frac{\partial P_i^*}{\partial s_j} = 2y_i(pr_i^* - c_i - \rho_i)\frac{\partial pr_i^*}{\partial s_j}. \tag{12}$$

From it we can infer that CP i's profit increases as a result of QoS level improvement by a competing CP j if and only if the QoS level improvement results in an increase in CP i's price. This happens when P_i^* increases on the interval $[0, s_{ij}^0]$ and decreases on the remaining interval $(s_{ij}^0, \overline{rt}]$. In regard to profit variation trends, on its own QoS level improvement, a dominant trend for a CP is not observed. However, we make two observations based on the holding of the following relationship

$$\frac{\partial P_i^*}{\partial s_j} = 2y_i(pr_i^* - c_i - \rho_i)\frac{\partial pr_i^*}{\partial s_j} - \frac{\rho_i}{(\overline{rt} - s_i)^2}. \tag{13}$$

If a CP i increases its QoS level from 0 to a positive value and this results in its price decrease, i's equilibrium profits become a decreasing function of its QoS level at all times. Thus, in such a case i is better off providing minimal QoS level to its customers. However, when CP i's QoS level increases from 0 to a positive value resulting in an increase in its price charged to customers, there exists a QoS level s_i^b such that the equilibrium profit alternates arbitrarily between increasing and decreasing in the interval $[0, s_i^b)$, and decreases when $s_i \geq s_i^b$.

Convergence to Nash equilibria: Since the price game in question has a unique and optimal Nash equilibria, it can be easily found by solving the system of first order conditions, $\frac{\partial P_i}{\partial pr_i} = 0$ for all i.

Remark. It is true that the existence of NE in convex games is not surprising in view of the general theory, but what is more important is whether a realistic modeling of our problem at hand results in a convex game. Once we can establish that our model results in a convex game, we have a straightforward result of the existence of NE from the game theory literature. This is exactly what we do in the paper, i.e., to show that our model is realistic and indeed leads to a convex game thus leading further to the existence of NE.

4. Game 2—price–QoS game

In this section, we analyze the game in which the CPs compete for both, price as well as QoS levels. In the process of analyzing Game 2, we also derive Game 3, as a special case of Game 2, and state results pertaining to Game 3.

Game description

Players: individual cloud providers; *Game type:* non-cooperative, i.e., no interaction between CPs; *Strategy space:* price in range $[pr_i^{min}, pr_i^{max}]$ and QoS level s_i; *Player goal:* to maximize its individual utility $U_i = P_i$.

We have the following theorem regarding equilibrium results.

Theorem 2. *Let* $\overline{rt} \leq \sqrt[3]{\frac{4\underline{y}\underline{\rho}}{(\overline{x'})^2}}$, *where* $\underline{y} = min_i\, y_i$, $\underline{\rho} = min_i\, \rho_i$, $\overline{x'} = max_i\, x_i'(0)$. *There exists a Nash equilibrium* (pr^*, s^*), *which satisfies the following system of equations:*

$$\frac{\partial P_i}{\partial pr_i} = -y_i(pr_i - c_i - \rho_i) + \lambda_i = 0, \quad \forall i, \tag{14}$$

and satisfies the condition that either $s_i(pr_i)$ *is the unique root of* $x_i'(s_i)(pr_i - c_i - \rho_i) = \frac{\rho_i}{(\overline{rt} - s_i)^2}$ *if* $pr_i \geq c_i + \rho_i(1 + \frac{1}{\overline{rt}^2 x_i'(0)})$ *or* $s_i(pr_i) = 0$ *otherwise. Conversely, any solution of these two equations is a Nash equilibrium.*

Sensitivity analysis: We know that $s_i(pr_i)$ depends on $x_i'(s_i)$ and pr_i. Thus, from the *implicit function theorem* [11] we infer that the QoS level of CP i increases with the increase in its Nash equilibrium price. We have the following relationship for $pr_i > c_i + \rho_i(1 + \frac{1}{\overline{rt}^2 x_i'(0)})$,

$$s_i'(pr_i) = \frac{x_i'(s_i)}{x_i''(s_i)(pr_i - c_i - \rho_i) - \frac{\rho_i}{(\overline{rt} - s_i)^2}} > 0, \tag{15}$$

whereas $s_i'(pr_i) = 0$ for $pr_i < c_i + \rho_i(1 + \frac{1}{\overline{rt}^2 x_i'(0)})$. We also notice that for $pr_i > c_i + \rho_i(1 + \frac{1}{\overline{rt}^2 x_i'(0)})$, s_i^* increases concavely with pr_i^*. The value of $s_i(p_i)$ obtained from the solution of the equation $x_i'(s_i)(pr_i - c_i - \rho_i) = \frac{\rho_i}{(\overline{rt} - s_i)^2}$ if $pr_i \geq c_i + \rho_i(1 + \frac{1}{\overline{rt}^2 x_i'(0)})$, can be fed into Eq. (15) to compute the price vector. The system of equations that result after substitution is *non-linear* in vector **pr** and could have multiple solutions, i.e., multiple Nash equilibria.

Inferences from sensitivity analysis: Games 1, 2, and 3 give us non-intuitive insights to the price–QoS changes by individual CPs. We observe that the obvious intuitions of equilibrium price decrease of competing CPs with increasing QoS levels and vice-versa do not hold under all situations and sensitivity analysis provide the conditions under which the counter-result holds. Thus, the intricate nature of non-cooperative strategy selection by individual CPs and the interdependencies of individual strategies on the cloud market make cloud economics problems interesting.

Convergence to Nash equilibria: Since multiple Nash equilibria might exist for the price vectors for the simultaneous price–QoS game, the *tatonnement scheme* [9,12] can be used to prove convergence. This scheme is an iterative procedure that numerically verifies whether multiple price equilibria exist, and uniqueness is guaranteed if and only if the procedure converges to the same limit when initial values are set at pr^{min} or pr^{max}. Once the equilibrium price vectors are determined, the equilibrium service levels are easily computed. If multiple equilibria exist the cloud providers select the price equilibria that is component-wise the largest.

Regarding the case when CP price vector is given, we have the following corollary from the result of Theorem 2, which leads us to equilibrium results of *Game 3*, a special case of Game 2.

Corollary 2. *Given any CP price vector, pr^f, the Nash equilibrium $s(pr^f)$ is the dominant solution in the QoS level game between CPs, i.e., a CP's equilibrium QoS level is independent of any of its competitors cost or demand characteristics and prices. When $s_i(pr^f) > 0$, the equilibrium QoS level is increasing and concave in pr_i^f, with $s_i'(pr_i^f) = \frac{-x_i'(s_i)}{x_i''(s_i)(pr_i^f - c_i - \rho_i) - \frac{2\rho_i}{(\overline{rt} - s_i)^3}}$.*

We observe that Game 3 being a special case of Game 2 entails a unique Nash equilibrium, whereas Game 2 entails multiple Nash equilibria.

5. Optimization framework for capacity provisioning

In this section, we develop optimization models for optimally provisioning capacity in both, single-tier as well as multi-tier cloud networks. As mentioned in previous sections, the term 'capacity' has a queueing-theoretic notion to it and is the service rate of a queueing system processing user requests, i.e., it is the number of user requests processed per unit of time. The capacity measure can be translated to allocating hardware and other system resources optimally so as to satisfy user QoS demands. In the following subsections, we first deal with the capacity analysis in single tier clouds, which is followed by the analysis in multi-tier cloud networks.

5.1. Single-tier case

We model each CP i as an M/M/1 queueing system with first-come, first-serve (FCFS) scheduling, where end-user requests arrive as a Poisson process with mean rate λ_i, and gets serviced at a rate μ_i. We adopt an M/M/1 queueing system because of three reasons: (1) queueing theory has been traditionally used in request arrival and service problems, (2) for our problem, assuming an M/M/1 queueing system ensures tractable analyses procedures that entails deriving nice closed form expressions and helps understand system insights in a non-complex manner, without sacrificing a great deal in capturing the real dynamics of the actual arrival–departure process, and (3) the Markovian nature of the service process helps us generalize expected steady state analysis and percentile analysis together. We assume that each CP adopts the FCFS scheduling policy because they serve a single class of end-users with the same QoS level guarantees.

The metric for end-user satisfaction in queueing systems is response/waiting time. The response time rt_i may be defined, either in terms of the expected steady state response time, i.e., $rt_i = E(RT_i)$, or in terms of ϕ-percentile performance, $rt_i(\phi)$, where $0 < \phi < 1$. Thus, in terms of ϕ-percentile performance,[10] $P(RT_i < rt_i(\phi)) = \phi$. According to the theory of M/M/1 queues, we have the following standard results [17].

$$rt_i = \frac{1}{\mu_i - \lambda_i}, \tag{16}$$

$$rt_i(\phi) = \frac{ln(\frac{1}{1-\phi})}{\mu_i(\phi) - \lambda_i}, \tag{17}$$

$$\mu_i = \lambda_i + \frac{1}{rt_i}, \tag{18}$$

and

$$\mu_i(\phi) = \lambda_i + \frac{ln(\frac{1}{1-\phi})}{rt_i(\phi)}. \tag{19}$$

[10] As an example, in cloud networks we often associate provisioning power according the 95th percentile use. Likewise, we could also provision service capacity by accounting for percentile response time guarantees.

Eqs. (20) and (22) follow from the fact that for M/M/1 queues, the following result holds,

$$P(RT_i < rt_i(\phi)) = \phi = 1 - e^{-(\mu_i - \lambda_i rt_i(\phi))}. \tag{20}$$

The inverse of $rt_i(rt_i(\phi))$ is $s_i(s_i(\phi))$, which is the advertised QoS level guarantee of CP i to its end-users. Thus, we observe from Eqs. (21) and (22) that the queueing service rate (capacity) is *linear* in λ_i and $s_i(s_i(\phi))$. Since C_i is proportional to $\mu_i(\mu_i(\phi))$, we infer that C_i is linear in λ_i and $s_i(q_i(\phi))$. *Our aim in this paper is to find the optimal $\mu i(\mu_i(\phi))$ for each CP i such that its advertised QoS level guarantees to its end-users are satisfied, without wasting any resources.*

Assuming that it takes a cost of ρ_i for CP i to provision a single unit of service capacity, we have the following optimization problems considering the expected value and percentile value of response time respectively.

$$min\ \rho_i \mu_i$$

subject to

$$\frac{1}{\mu_i - \lambda_i} \leq rt_i \quad \forall i$$

and

$$min\ \rho_i \mu_i(\phi)$$

subject to

$$\frac{log(\frac{1}{1-\phi})}{\mu_i(\phi) - \lambda_i} \leq rt_i(\phi) \quad \forall i.$$

5.2. Multi-tier case

In order to model the multi-tier case, we model a given cloud network for CP i as a network of queues. Each queue in the network acts as an M/M/1 queue serving end-user requests in an FCFS manner. We assume that the queueing network is an *open Jackson network* [17]. We also assume the queueing network for any CP is distinct from other CP queuing networks, i.e., for CP i, there is no queue in its network that serves any other CP j, $\forall j \neq i$. Each queue is representative of a *tier* in a cloud network and is represented as a vertex/node in the open Jackson network. The departure process of one tier/level is an arrival process for the next tier. We define the following notations in relation to our analysis of queueing networks for CP i

V^i - set of n vertices in an open Jackson network for CP i.
π_j^i - fraction of end user requests that start servicing at node j.
p_{jk}^i - probability that a user request moves to node k after getting service from node j.
P^i - matrix of p_{jk}^i values and is sub-stochastic in nature, i.e., $Lt_{n\to\infty}(P^i)^n = 0$
μ_j^i - service rate of node $j \epsilon V^i$
ρ_j^i - capacity cost per unit of service rate at node j.
Ω^i - vector of aggregate arrival rates for CP i.

The vector of arrival rates for each CP i is expressed as recursive expression of the form

$$\Omega^i = \lambda_i \pi^i + (P^i)^T \Omega^i. \tag{21}$$

Solving the above equation, we get

$$\Omega^i = \lambda_i \delta^i, \tag{22}$$

where the vector $\delta^i = (I - (P^i)^T)^{-1}\pi^i$. According to queueing theory results regarding networks of queues, we get the following for expressions for each CP i *(for the expected value case of response time)*[11]

$$E[requests\ at\ node\ j] = \frac{\Omega_j^i}{\mu_j^i - \Omega_j^i} \tag{23}$$

and

$$E[total\ number\ system\ requests] = \lambda_i \sum_{j \epsilon V^i} \frac{\delta_j^i}{\mu_j^i - \Omega_j^i}. \tag{24}$$

[11] Due to the Markovian nature of the service process, the case for general percentiles is exactly similar to the case for expected steady state analysis. The expressions remain nearly the same apart from a constant factor multiplication.

By Little's law, we have

$$s_i^{-1} = \sum_{j \epsilon V^i} \frac{\delta_j^i}{\mu_j^i - \lambda_i \Omega_j^i}. \tag{25}$$

We now prove through the following theorem that even in multi-tier cloud networks, C_i is linear in λ_i and s_i, for each CP i. This fact regarding linearity is *important* when it comes to the case of analyzing price–QoS games.

Theorem 3. *The capacity provisioning cost, C, for each cloud provider in a multi-tier cloud network is linear in their user arrival rate and the advertised QoS level guarantee.*

Proof. Each cloud provider i is willing to minimize their capacity costs. Thus it selects $\mu^i = (\mu_j^i : j \epsilon V^i)$ such that it is the solution of the following constrained optimization problem

$$min \sum_{j \epsilon V^i} \rho_j^i \mu_j^i$$

subject to

$$\sum_{j \epsilon V^i} \frac{\delta_j^i}{\mu_j^i - \lambda_i \delta_j^i} \leq s_i^{-1}.$$

Applying Karush–Kuhn–Tucker (KKT) conditions [25] for optimality, we have

$$\rho_j^i = \frac{\gamma \delta_j^i}{(\mu_{j(opt)}^i - \lambda_i \delta_j^i)^2}, \quad j \epsilon V^i, \tag{26}$$

where γ is the Lagrange multiplier. From the previous equation we get

$$\mu_{j(opt)}^i - \lambda_i \delta_j^i = \sqrt{\gamma} \sqrt{\frac{\delta_j^i}{\rho_j^i}}, \quad j \epsilon V^i. \tag{27}$$

The minimum cost of CP i evaluates to $\sum_{j \epsilon V^i} \rho_j^i \mu_{j(opt)}^i$, which is of the form $A_1 \lambda_i + A_2 s_i$, where

$$A1 = \sum_{j \epsilon V^i} \rho_j^i \delta_j^i \tag{28}$$

and

$$A2 = \left(\sum_{j \epsilon V^i} \sqrt{\delta_j^i \rho_j^i} \right)^2. \tag{29}$$

Thus, the capacity provisioning cost per CP in a multi-tier cloud network is linear in their user arrival rate and the advertised QoS level guarantee. We emphasize that the theorem holds *(due to the Markovian nature of the service times)* when we consider the response-time as a percentile parameter, rather than an expected value. □

Optimization Problems: We have the following two optimization problems for multi-tier networks considering the expected value and percentile value of response time respectively.

$$min \sum_{j \epsilon V^i} \rho_j^i \mu_j^i$$

subject to

$$\sum_{j \epsilon V^i} \frac{\delta_j^i}{\mu_j^i - \lambda_i \delta_j^i} \leq s_i^{-1}$$

and

$$min \sum_{j \epsilon V^i} \rho_j^i \mu_j^i(\phi)$$

subject to

$$\sum_{j \epsilon V^i} log\left(\frac{1}{1-\phi}\right) \frac{\delta_j^i}{\mu_j^i(\phi) - \lambda_i \delta_j^i} \leq s_i^{-1}(\phi).$$

The optimization problems for the single-tier and multi-tier cases provide a framework via which resources can be provisioned in the cloud in a manner so as to minimize over-provisioning in a dynamic manner.

6. Conclusion and future work

In the first part of the paper, we developed inter-organizational economic models for pricing cloud network services when several cloud providers co-exist in a market, servicing a single application type. We devised and analyzed three price–QoS game-theoretic models relevant to cloud networks. We proved that a *unique* pure strategy Nash equilibrium (NE) exists in two of our three QoS-driven pricing models. In addition, we also showed that player dynamics converge to NE's *converge*; i.e., there is a *practically implementable* algorithm for each model that computes the NE/s for the corresponding model. Thus, even if no unique Nash equilibrium exists in some of the models, we are guaranteed to find the largest equilibria (preferred by the CPs) through our algorithm. Regarding convergence to Nash equilibria, it is true that it could take a long time for convergence of Nash equilibria (computing NE is PPAD Complete [18]), however in 95% of the cases in practical economic markets, NE is achieved in a decent amount of time.

Our price–QoS models can drive optimal resource provisioning in cloud networks. The NE price and QoS levels for each cloud provider drive optimal end-user demand in a given time period w.r.t. maximizing individual CP profits under competition. Servicing end-user demands requires provisioning capacity. Once the optimal values are computed, the power of virtualization in cloud networks makes it possible to execute dynamic resource provisioning in a fast and efficient manner in multiple time periods. In this regard, in the second part of the paper, we developed an optimization framework for single-tiered and multi-tiered cloud networks to compute the optimal provisioned capacity once the equilibrium price and QoS levels for each CP have been determined. As part of future work, we plan to extend our analysis to the case where cloud providers are in simultaneous competition with other CPs on *multiple* application types.

Appendix

Proof of Theorem 1. For a given service level vector $\mathbf{s}$, each CP i reserves a capacity of $\frac{1}{rt_i} = \frac{1}{\overrightarrow{rt} - s_i}$. Consider the game G with profit/utility functions for each CP i represented as

$$P_i = x_i(s_i) - y_i p_i - \sum_{j \neq i} \alpha_{ij}(s_j) + \sum_{j \neq i} (\beta_{ij} p_j)(pr_i - c_i - \rho_i) - W, \tag{30}$$

where

$$W = \frac{\rho_i}{\overrightarrow{rt} - s_i}.$$

Since $\frac{\partial^2 P_i}{\partial pr_i \partial pr_j} = \beta_{ij}$, the function P_i is *supermodular*.[12] The strategy set of each CP i lies inside a closed interval and is bounded, i.e., the strategy set is $[pr_i^{min}, pr_i^{max}]$, which is a *compact* set. Thus, the pricing game between CPs is a *supermodular game* and possesses a Nash equilibrium [13]. Since $y_i > \sum_{j \neq i} \beta_{ij},\ i = 1, \ldots, n$ (by Eq. (7)), $-\frac{\partial^2 P_i}{\partial pr_i^2} > \sum_{i \neq j} \frac{\partial^2 P_i}{\partial pr_i \partial pr_j}$ and thus the Nash equilibrium is unique. Rewriting Eq. (11) and using Eq. (6), we get $\lambda_i^* = y_i(pr_i^* - c_i - \rho_i)$. Substituting λ_i^* in Eq. (9), we get $P_i^* = y_i(pr_i^* - c_i - \rho_i)^2 - \frac{\rho_i}{\overrightarrow{rt} - s_i}$ □

Proof of Corollary 1. Since the inverse of matrix $\mathbf{M}$, i.e., M^{-1} exists and is greater than or equal to 0 [14], from $pr^* = M^{-1}(\overline{\mathbf{x}}(\mathbf{s}) + \mathbf{z})$ (Eq. (11)), we have pr_i^* is increasing in $\{c_i, \rho_i\, i = 1, 2, \ldots, n\}$. Again, from Lemma 2 in [14], we have $\delta_i \equiv y_i(M^{-1})_{ii} \Rightarrow 0.5 \leq \delta_i < 1$, where δ_i is the degree of *positive externality*[13] faced by CP i from other CP (price, QoS) parameters, and it increases with the β coefficients. This leads us to $\frac{\partial pr_i}{\partial c_i} = \frac{\partial pr_i}{\partial \rho_i} = y_i(M^{-1})_{ii} = \delta_i > 0$. Therefore, we show in another different way that pr^* is increasing in $\{c_i, \rho_i,\ i = 1, 2, \ldots, n\}$. Since M^{-1} exists and is greater than or equal to 0, we again have $\frac{\partial \lambda_i}{\partial c_i} = \frac{\partial \lambda_i}{\partial \rho_i} = y_i(\frac{\partial pr_i}{\partial \rho_i} - 1) = y_i(\frac{\partial pr_i}{\partial c_i} - 1) = y_i(\delta_i - 1) < 0$, from which we conclude that λ^* is decreasing in $\{c_i, \rho_i,\ i = 1, 2, \ldots, n\}$. Part (b) of the corollary directly follows from the fact that the inverse of matrix $\mathbf{M}$, i.e., M^{-1} exists, is greater than or equal to 0, and every entry of M^{-1} is increasing in β_{ij} coefficients. □

Proof of Theorem 2. To prove our theorem, we just need to show that the profit function P_i is *jointly concave* in (pr_i, s_i). Then by the *Nash–Debreu* theorem [15], we could infer the existence of a Nash equilibria. We know the following results for all CP i

$$\frac{\partial P_i}{\partial pr_i} = -y_i(pr_i - c_i - \rho_i) + \lambda_i \tag{31}$$

[12] A function $f : R^n \to R$ is supermodular if it has the following increasing difference property, i.e., $f(m_i^1, m_{-i}) - f(m_i^2, m_{-i})$, increases in m_i for all $m_i^1 > m_i^2$ in (pr_i, pr_j). The readers are referred to [16] for more details on supermodularity.

[13] A positive externality is an external benefit on a user not directly involved in a transaction. In our case, a transaction refers to a CP setting its price and QoS parameters.

and

$$\frac{\partial P_i}{\partial \theta_i} = x_i'(s_i)(pr_i - c_i - \rho_i) - \frac{\rho_i}{(\overline{rt} - s_i)^2}. \tag{32}$$

Thus, $\frac{\partial^2 P_i}{\partial pr_i^2} = -2y_i < 0$, $\frac{\partial^2 P_i}{\partial s_i^2} = x_i''(s_i)(pr_i - c_i - \rho_i) - \frac{2\rho_i}{(\overline{rt}-s_i)^3} < 0$, $\frac{\partial^2 P_i}{\partial s_i \partial pr_i} = x_i'(s_i)$. We determine the determinant of the Hessian as $-2y_i(x_i''(s_i)(pr_i - c_i - \rho_i) - \frac{\rho_i}{(\overline{rt}-s_i)^2} \geq 0$ (the sufficient condition for P_i to be jointly concave in (pr_i, s_i)), if the following condition holds:

$$\frac{4y_i\rho_i}{\partial pr_i^2} \geq (x_i'(s_i))^2 \Leftrightarrow \overline{rt} \leq min_{s_i} \sqrt[3]{\frac{4y_i\rho_i}{(x_i'(s_i))^2}} = \sqrt[3]{\frac{4y_i\rho_i}{(x_i'(0))^2}}, \tag{33}$$

where the last equality follows from the fact that $x_i' > 0$ and x_i' is decreasing. Now since $pr^* = pr^*(s^*)$, by Theorem 1 it is in the closed and bounded interval $[pr^{min}, pr^{max}]$ and must therefore satisfy Eq. (15). Again from Eq. (31), we have $\frac{\partial P_i}{\partial s_i} \to -\infty$ as s_i tends to $\overline{rt}$, which leads us to the conclusion that $s_i(pr_i)$ is the unique root of $x_i'(s_i)(pr_i - c_i - \rho_i) = \frac{\rho_i}{(\overline{rt}-s_i)^2}$ if $pr_i \geq c_i + \rho_i(1 + \frac{1}{\overline{rt}^2 x_i'(0)})$ or $s_i(pr_i) = 0$ otherwise. □

Proof of Corollary 2. *Substituting $pr^{max} = pr^{min} = pr^f$ into Theorem 2 leads us to the fact that $s(pr^f)$ is a Nash equilibrium of the QoS level competition game amongst CPs and that it is also a* unique *and a* dominant *solution, since $s(pr^f)$ is a function of pr_i, c_i, and ρ_i only. (Following from the fact that $s_i(pr_i)$ is the unique root of $x_i'(s_i)(pr_i - c_i - \rho_i) = \frac{\rho_i}{(\overline{rt}-s_i)^2}$ if $pr_i \geq c_i + \rho_i(1 + \frac{1}{\overline{rt}^2 x_i'(0)})$ or $s_i(pr_i) = 0$ otherwise.)* □

References

[1] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, M. Zaharia, Above the clouds: a Berkeley view of cloud computing, Technical Report, EECS, U. C. Berkeley, 2009.
[2] S.C.M. Lee, J.C.S. Lui, On the interaction and competition among Internet service providers, IEEE Journal on Selected Areas in Communications 26 (2008).
[3] S. Shakkotai, R. Srikant, Economics of network pricing with multiple ISPs, IEEE/ACM Transactions on Networking 14 (2006).
[4] P. Hande, M. Chiang, R. Calderbank, S. Rangan, Network pricing and rate allocation with content-provider participation, in: IEEE INFOCOM, 2010.
[5] L. Jiang, S. Parekh, J. Walrand, Time-dependent network pricing and bandwidth trading, in: IEEE BoD, 2008.
[6] J.K. Mackie-Mason, H.R. Varian, Pricing congestible network resources, IEEE Journal on Selected Areas in Communications 13 (1995).
[7] D. Songhurst, F. Kelly, Charging schemes for multiservice networks, in: 15th International Teletra?c Congress, 1997.
[8] P. Dube, R. Jain, C. Touati, An analysis of pricing competition for queued services with multiple providers, in: ITA Workshop, 2008.
[9] H.R. Varian, Microeconomic Analysis, Norton, 1992.
[10] M.E. Wetzstein, Microeconomic Theory: Concepts and Connections, South Western, 2004.
[11] W. Rudin, Principles of Mathematical Analysis, Mc.Graw Hill, 1976.
[12] K. Arrow, Handbook of Mathematical Economics, North Holland, 1981.
[13] X. Vives, Nash equilibrium and strategic complementarities, Journal of Mathematical Economics 19 (1990).
[14] F. Bernstein, A. Federgruen, Comparitive statics, strategic complements, and substitutes in oligopolies, Journal of Mathematical Economics 40 (2004).
[15] D. Fudenberg, J. Tirole, Game Theory, MIT Press, 1991.
[16] D.M. Topkis, Supermodularity and Complementarity, Princeton University.
[17] D. Bertsekas, R. Gallager, Data Networks, Prentice Hall Inc, 1988.
[18] C. Daskalakis, P.W. Goldberg, C.H. Papadimitrou, The complexity of computing a Nash equilibrium, SIAM Journal of Computing 39 (1) (2009).
[19] D. Gmach, J. Rolia, L. Cherkasova, A. Kemper, Capacity management and demand prediction for next generation data centers, in: IEEE International Conference on Web Services, 2007.
[20] T. Wood, L. Cherkasova, K. Ozonat, P. Shenoy, Profiling and modeling resource usage of virtualized applications, in: ACM International Conference on Middleware, 2008.
[21] D. Kusic, N. Kandasamy, Risk-aware limited lookahead control for dynamic resource provisioning in enterprise computing systems, in: IEEE ICAC, 2006.
[22] P. Padala, K.G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, A. Merchant, K. Salem, Adaptive control of virtualized resources in utility computing environments, in: ACM SIGOPS, 2007.
[23] X. Meng, C. Isci, J. Kephart, L. Zhang, E. Bouilett, Efficient resource provisioning in compute clouds via VM multiplexing, in: ACM ICAC, 2010.
[24] J. Dejun, G. Pierre, C-H. Chi, Autonomous resource provisioning for multi-service web applications, in: ACM WWW, 2010.
[25] S. Boyd, L. Vanderberghe, Convex Optimization, Cambridge University Press, 2005.