Revenue Maximization for Cloud Computing Services

Cinar Kilcioglu, Costis Maglaras
Graduate School of Business, Columbia University, New York, NY 10027
ckilcioglu16@gsb.columbia.edu, c.maglaras@gsb.columbia.edu

We study a stylized revenue maximization problem for a provider of cloud computing services, where the service provider (SP) operates an infinite capacity system in a market with heterogeneous customers with respect to their valuation and congestion sensitivity. The SP offers two service options: one with guaranteed service availability, and one where users bid for resource availability and only the "winning" bids at any point in time get access to the service. We show that even though capacity is unlimited, in several settings, depending on the relation between valuation and congestion sensitivity, the revenue maximizing service provider will choose to make the spot service option stochastically unavailable. This form of intentional service degradation is optimal in settings where user valuation per unit time increases sub-linearly with respect to their congestion sensitivity (i.e., their disutility per unit time when the service is unavailable) – this is a form of "damaged goods." We provide some data evidence based on the analysis of price traces from the biggest cloud service provider, Amazon Web Services.

Key words: congestion, revenue management, service level, damaged goods

1. Introduction

Data made available in social networks, media and entertainment, electronic commerce, and mobile is exploding. Firms across industries are increasingly focusing on the use of data analytics to generate insightful and actionable insights to improve their profitability and growth, improve customer experience, design new and better products and services. Together these trends have led to a significant increase in IT storage and computing requirements across industries, and apart from significant infrastructure investments in computing and data storage clusters, they have led to increased support, management and maintenance costs. The operating loads of these large corporate storage and computing clusters exhibit significant intraday and seasonal variability, and additionally firms want flexibility for rapid growth in resource requirements as their needs evolve and mature. In this environment, cloud computing –a form of outsourcing of the aforementioned physical IT infrastructure resources– has become a cost effective alternative for these firms.

In broad terms, cloud computing refers to the provision of computing resources, such as storage, data management, and processing, over a network of remote servers hosted on a

remote data center location and accessible via the internet, which is broadly available and at abundant speeds. Currently, Amazon, Google, and Microsoft are the leading providers of cloud computing services to a variety of customers, ranging from individuals and small firms, to global media companies and government agencies. These customers differ with respect to their resource needs, duration, valuation and sensitivity to service level. For instance, while a researcher who does not have a strict time constraint and has a limited budget may prefer to procure computing power anytime within a week and pay little. On the other hand, an online retailer that hosts its web servers in the cloud is very sensitive to service availability and the quality (speed) of the rendered service. This heterogeneity with respect to price and congestion sensitivity allows service providers to offer a menu of product options to segment and better serve this market, essentially offering hosted computing resources at different price levels depending on their anticipated service availability (e.g., as measured by the % of time that the resource will be available).

This paper studies a problem of market segmentation for a revenue maximizing (monopolist) service provider (SP) of cloud computing resources that offers two classes of service: guaranteed (on-demand instances) and best effort (spot instances). The latter is procured via a second price auction. This problem is motivated by the service menu offered by Amazon Web Services (AWS), the largest SP in the market currently. Insights extracted from asymptotic analysis of large scale multi-server systems suggest that the observed variation in spot prices is not consistent with the natural stochastic fluctuations between a two-class priority service system. Moreover, it is typically believed that these SP's are not capacity constrained in this stage, but rather experiencing a rapid phase of infrastructure investment aiming to capture market share. Motivated by these observations, we study a SP that operates a system with infinite capacity, and note that under that assumption there is no competition for scarce resources between the two service classes or amongst the users bidding for spot service; specifically all users bidding higher than the SP's reserve price get access to uninterrupted service. The quality of a product is defined as the fraction of time the product is available to customers. While guaranteed service offers 100% availability with a fixed price, the quality level and the payment depend on customers' bids in best effort service. Each customer gains some positive utility from the service proportional to the time that the service is received and suffers a negative utility proportional to the length of time that the service is unavailable. The market is heterogeneous, and, specifically, users

differ with respect to two parameters: valuation per hour of service and disutility per hour of service disruption (sensitivity to congestion). The former is how much customers are willing to pay for one hour service, and the latter is how much disutility one hour of service interruption creates. For example, for an online retailer hosting its web servers, valuation is the customer's willingness to pay to have the web server running for one hour, and sensitivity to congestion is the cost of not having the server running, which may include the lost revenue or profit margin as well as lost goodwill from affected customers. Both valuation and sensitivity to congestion are private information and thus unknown to the SP. All users are assumed to have infinite duration service requirements.

We formulate and solve the deterministic SP's revenue maximization problem. We treat two cases separately. First, we study the case where valuation per unit time grows sub-linearly as a function of the disutility per unit time of service disruption, i.e., where `(valuation/time)/(cong. sens./time)` $\downarrow$ as `(cong. sens./time)` $\uparrow$. In other words, for users that are congestion sensitive, disutility due to service disruption grows faster than the user's valuation. We assume that the user's valuation per unit time of service is an affine, increasing, function of her congestion cost per unit time of service disruption. In this case user types are one-dimensional, and we assume that user congestion costs per unit time are independent identically distributed (i.i.d.) draws from a continuous distribution, with a strictly positive density function and bounded support. We model the prevailing spot price as a discrete process (e.g., in \$.01 increments per hour) and focus on the associated steady state distribution. We assume that the SP can select the steady state distribution, i.e., the long run average fraction of time for which the spot price spends at each price level; if the SP's reserve price is constant through time, then the steady state distribution will reduce to a point mass at that respective level. Users (have infinite service time requirements) observe the steady state distribution of the spot price path, and choose between guaranteed and spot, and, if they select the spot service option, they also determine how much to bid. We prove that i) the SP should set the price levels of the spot service option such that the lowest spot price level will be below the lowest valuation across all users in the market (that is, nobody is priced out); ii) it is optimal to use two distinct price levels in spot service for positive fractions of time, respectively, and offering more than two price levels does not generate more revenue for the SP; and iii) the fraction of time that the spot service price is "high" depends on the coefficients of the affine relation between

congestion cost rate and valuation rate, but not the distribution of types itself. Finally, if the valuation rate grows sub-linearly with respect to the congestion cost rate but the respective relation is general (not affine), then we show that it may be optimal for the SP to offer multiple (>2) price levels for the spot service option.

The second case we study is one where the valuation rate increases faster than the user's congestion cost rate, in this case we prove that it is never optimal to offer spot service. Intuitively, in this setting congestion sensitive users are willing to pay increasingly high amounts, and the SP is not willing to sacrifice any revenue from these high types by offering an incentive compatible lower priced spot price option. Therefore, if more congestion sensitive customers have comparatively higher valuations, then it is optimal to serve only the high-valuation market segment by offering the high quality service.

Lastly, we analyze the price traces of over 1,000 products that the dominant provider in the market offers for a six-month period, and present descriptive statistics that sheds light on the dynamics of the spot price. Calibrating our model on the observed data, we offer some insight on the dynamics of spot price valuations, and characterize the relative magnitude of valuation rate to the congestion rate; the latter may be as much as 10 times larger than the former.

The remainder of this paper is structured as follows. This section concludes with a brief literature survey. Section 2 offers a short introduction to the services and pricing that we encounter in today's cloud computing SPs. Section 3 describes our model, which we analyze in Sections 4 and 5 that are organized with respect to the relation of user valuation and user congestion sensitivity per unit time. Finally, Section 6 offers a more detailed look into the pricing data from the currently largest provider of cloud computing, Amazon Web Services, and briefly discuss some of its implications.

Our work is related to the literature on "economics of queues," which goes back to the work of Naor (1969) that introduced the study of strategic customer behavior in a queueing setting. Mendelson (1985) and Mendelson and Whang (1990) studied (primarily) social welfare optimization in an $M/M/1$ system serving a market of heterogeneous, utility maximizing customers. Afèche (2013) studied the revenue maximization problem for a SP operating in a market with two segments that differ with respect to their delay sensitivity, and importantly showed that the SP may use the notion of "strategic delay" to optimally segment the market and optimize the system's revenue rate. Strategic delay

amounts to (artificially) increasing the realized waiting time of some customers beyond the waiting time that they would experience due to the system's congestion effects. This is akin to the idea of "damaged goods" introduced earlier on in economics and marketing, e.g., Deneckere and McAfee (1996) and McAfee (2007) that showed that profit maximizing firms may intentionally "crimp" their products to price discriminate, and Pareto improve performance; these papers provide striking examples from high-tech industry; see also, Anderson and Dana (2009).

Our model does not involve any congestion phenomena that arise due to the dynamics of a finite capacity physical system, and as such resembles in its nature the marketing and economics references on damaged goods. In terms of model formulation and motivation, however, it is closer to several papers from the economics of queues literature that we highlight below. Afèche and Pavlin (2015) studied a model with one-dimensional types, where the valuation is a linear function of the delay cost parameter. For this model they characterized for a SP that operates an $M/M/1$ system. We will consider the same model in §4.1 and study the SP's revenue maximizing solution in that case. Our model differs from the one above in its utility function: specifically, users extract value from the service and pay only when service is available, and incur disutility but stop paying when service is interrupted. Our result that shows that the use of "damaged goods" may be optimal is similar to theirs. The affine model is an example of a model where valuation grows sub-linearly as a function of the congestion sensitivity. §4.2 shows that when the monotonicity result holds but the relation between the two parameters is general, then the optimal solution may involve again the use of damaged goods but the structure of the optimal policy is more complex. §5 looks at the case where the valuation rate grows super-linearly as a function of the congestion cost parameter, which is akin to the model studied in Katta and Sethuraman (2005). Our utility function is again different and the details of the analysis are not the same, but one of the key findings that the use of damaged goods is not needed is consistent with their results (considered when capacity grows large and the system becomes uncongested). Nazerzadeh and Randhawa (2015) look at a similar model as the one studied in Katta and Sethuraman (2005) and among other things show that in the unconstrained capacity setting, offering one service level performs "well," which is consistent with our findings.

Our work is also related to the stream of work that studies economic optimization problems in a queue in the context of large scale systems. Maglaras and Zeevi (2003) showed that in a single type market where demand is elastic, the revenue maximizing operating regime in an $M/M/C$ system where the system size C and the market potential grow large is the so-called heavy-traffic regime. Maglaras et al. (2015) extended this analysis to multiple types of customers, establishing again, under some conditions, the phenomenon of strategic delay mentioned above. Finally, our model operates as a two class priority system. The asymptotic behavior of such a system in a multi-server setting was studied in Maglaras and Zeevi (2005).

Abhishek et al. (2012) ask a question similar to ours and analyze the problem of the SP using tools from mechanism design to show that offering only high a quality (guaranteed service) product with a fixed price generates more revenue than offering both high and low quality products at the same time. This result is in contrast to our findings in §4, as well as those in Afèche and Pavlin (2015). In our model of §4.2, users with valuation v_i have congestion cost parameter κ_i (deterministic), whereas in Abhishek et al. (2012) such users may have a random congestion cost parameter with distribution F_i. If we approximate our model in their setting by letting the capacity grow large, and, more importantly, restrict the support of their congestion rate parameter to a narrow support (centered around κ_i), then one of the key conditions needed for their main finding no longer holds, therefore removing the apparent inconsistency. Afèche and Mendelson (2004) studied the revenue maximization problem in a queue with priority auctions and generalized delay cost structure. They show that in some cases, revenue maximizing uniform pricing provides no or only little surplus loss. Moreover, using priority auctions instead of uniform pricing yields lower prices and higher utilization in the system.

In a recent study, Mitra and Wang (2015) consider a monopoly broadband access internet service provider that offers a guaranteed service with a usage fee, and a best effort service free of charge. In profit maximization setting, they show why best effort service "harvests" possibly new guaranteed service clients; at its core lies a stylized model for the dynamics of adoption of new users (applications) that start as best effort users (subsidized) and then some of these transition to successful applications that switch to guaranteed service quality.

Armbrust et al. (2010) provide an overview of cloud computing from different perspectives including cloud computing economics. Xu and Li (2013) show that throttling the resource generates more revenue than uniform usage pricing and performance guarantees can be provided with an extra fee. In their model customers differ only with respect to their valuation per unit time and each customer is allowed to choose different number of resources. Borgs et al. (2014) study a multiperiod pricing problem where the SP offers a service with varying capacity in a market that customers are strategic and heterogeneous in their valuations, arrival and departure periods. They used the cloud computing market as an example of such a setting, and provided an efficient algorithm to find a dynamic pricing mechanism that satisfies service guarantees. Savin et al. (2005) look at the problem of capacity allocation of rental equipment used by two customer types, with stochastic rental period requirements. They formulate the problem as a queueing control problem and provide a heuristic control based on a fluid model approximation. Baron (2003) considers a system (similar to cloud computing) that the SP shares her computing resources. He presents token-bucket admission control and pricing schemes. In this work customers compete for the shared resource. Our paper provides descriptive statistics and some analysis on a rich data set from a leading SP. Similar datasets have been analyzed in different works to find possible explanations for the observed price fluctuations. Agmon Ben-Yehuda et al. (2011) draw the conclusion that the SP varies her reserve price over time. They empirically show that the spot prices seem to follow trends that show significant regularity when views under an appropriate prism, and could be the result of the SP's control of the reserve price.

2. Glimpse of Cloud Computing Market and Pricing Mechanisms

The two key participants in the market for cloud computing are users and providers. The users can be individuals or companies requiring temporary (short-term) or permanent (long-term) computing resources that can be reached over the internet. The providers are the operators of the cloud computing services. Currently there are many small and large SPs in the market, with Amazon, Google, and Microsoft being the leading providers. There are three main services that cloud providers offer: software-as-a-service (SaaS), platform-as-a-service (PaaS), and infrastructure-as-a-service (IaaS). In this paper we are focusing on IaaS service, where the product is defined as the bundle of a machine type, an operating system, and a location.

Each provider offers its products under one or multiple price models. The dominant provider in the market is Amazon and it offers the richest pricing options. Currently, Amazon rents out its computing resources under three different pricing models: pay-as-you-go (on-demand instances), pay-as-you go under contract (reserved instances), and second price auction (spot instances). On-demand and reserved instances offer guaranteed service and in the sequel, we will focus on a model with only 2 service options: guaranteed and best effort, which we refer to "on-demand" and "spot."

We are focusing on two pricing models: on-demand and spot. Each product has a fixed hourly price in the on-demand market and users continue paying this fixed rate as long as they use the service. Amazon has no control of ending a running service, while customers can end their service at any point in time with no penalty. The spot market has a more complicated pricing structure. For each product, Amazon sets a reserve price, possibly time-varying, and customers bid their maximum willingness-to-pay per hour for that product. The spot price at any time point is defined as the minimum bid accepted at that time, which in some cases may be the reserve. The spot price fluctuates over time in response to variations to the available capacity not utilized by the "guaranteed" instances rented by Amazon, and to the number of active spot customers and their corresponding bids. If the bid of a particular customer falls below the spot price, this customer is temporarily out of access to the cloud (priced out) until the spot price falls again at or below her bid.

The data on hand shows the the spot price exhibits significant fluctuations over time. They may be around one tenth of the corresponding on-demand prices; and, can and do fluctuate to up to five or ten times of the corresponding on-demand prices; interrupting service for many spot instance customers, resulting in some form of disutility. If customers in the spot market bid sufficiently high and continuously pay the prevailing spot price (even when it is above the price of on-demand), in the long run they will receive uninterrupted service. The corresponding time-average spot price is cheaper than the corresponding on-demand price for some of the products, but certainly not all. We present further descriptive statistics in Section 6.

Another choice customers make is whether to use cloud or in-house resources for their computing needs. Table 1 shows the configuration and prices for a product family (*m4* machine types) with Linux operating system residing in *us-west-2* (Oregon) region. Among these products, we analyze *m4.xlarge* machine more closely. Hourly on-demand price for

this product is \$0.254/hr, while it can go up to \$0.374 in other regions. This product was available both in spot and on-demand markets approximately in the last 80 days of our time window. Usage in this period cost \$486 in on-demand market, while it was between \$113 to \$207 (depending on the subregion selected) in spot market. One year of continuous usage of this product costs \$2,208 in on-demand market. As a comparison, a similar in-house server (HP ProLiant DL380 Gen9 - Xeon E5-2620V3 2.4 GHz - 16 GB, which has 6 cores) costs around the same to purchase without any peripheral costs for mounting, networking, etc.. However, if one wants to rent a product for long-term continuous usage, reserved instances offers much cheaper options. For instance, the same product can be rented by paying \$1,271 upfront for one year of usage (see Armbrust et al. (2010) for a more detailed cost analysis).

Table 1 Prices in on-demand and reserved markets for a group of products and their configurations

Machine name	# cores	# RAM	price/hr	on-demand price/year	reserved price/year
m4.large	2	8	\$0.126	\$1,103.76	\$635
m4.xlarge	4	16	\$0.252	\$2,207.52	\$1,271
m4.2xlarge	8	32	\$0.504	\$4,415.04	\$2,541
m4.4xlarge	16	64	\$1.008	\$8,830.08	\$5,082
m4.10xlarge	40	160	\$2.52	\$22,075.2	\$12,706

3. Model Formulation

3.1. Detour: Asymptotic Behavior of Large Scale Multi-Server Systems

We briefly discuss a system where the SP has a finite processing capacity C and offers two nonsubstitutable service classes: guaranteed-rate (G) service and best-effort (BE) service. In the former, customers receive a constant service rate as long as there is capacity and are blocked otherwise; in the latter, service rate is dependent on the number of customers in the whole system. G service has priority over BE. BE users get one unit of capacity, if this is available, or share the available capacity (not used by G users) equally if there are more BE users connected than the available number of servers, thus experiencing congestion. Customers arrive to the system according to independent Poisson processes and the service requirements are exponentially distributed. Maglaras and Zeevi (2005) studied this system and showed that when the system size grows large, the G class occupies

$\alpha C + B(t)\sqrt{C}$ servers, where $0 < \alpha < 1$ and $B(t)$ is standard Brownian motion, and the remaining capacity, $(1-\alpha)C - B(t)\sqrt{C}$, is available to BE service. A similar analysis could be carried through under the auction model for BE service. The important observation is that the variation in the available capacity for BE users will be second order, and this would result in fluctuations of the prevailing spot price that would also be second order (i.e., small). This prediction does not agree with what we observe in the data. This suggests that perhaps a different mechanism gives rise to the fluctuations to the spot price that may be exogenous to the capacity dynamics of the BE class, as defined crudely by their supply-demand imbalance.

The above discussion has three important caveats that are worth noting. First, the model assumes the same (or reasonably similar) service durations for both services. Second, the model assumes each customer has unit demand. If users may demand a random number of servers and this follows a heavy-tailed distribution, it may be possible to observe big price spikes. The observed frequency and duration of price spikes would require frequent, random arrival of users with unusually large capacity needs that are short-lived (which may be implausible). Last, the model assumes that in equilibrium, the fraction of the overall system capacity consumed by each of the two service classes are comparable (and first order). If BE service used a very small fraction of the total capacity and the overall system was heavily utilized, then significant spot price fluctuations could emerge; e.g., the BE usage is of order $\sqrt{C}$, which is the same as the order of magnitude of the G service class, thus resulting in fluctuations of BE available capacity that are of the same order as the overall capacity used by BE. Nevertheless, in this case the revenue generated from BE service would be insignificant, rendering the parameter regime less interesting. It also seems unlikely that the data centers of large-scale cloud computing SPs are operating at full utilization at this point in time of rapid expansion and effort to capture market share.

3.2. The Infinite Capacity Model

Motivated from the above we will model the market as follows. The SP has infinite capacity and operates multiple resources and offers a service (or a product) from two distinct channels: guaranteed (G, on-demand instances) and best effort (BE, spot instances). Customers differ with respect to their willingness-to-pay for a unit-time service, v, and their congestion sensitivity parameter, κ. We assume that each customer has unit demand and infinite

service time. Customers are individual utility maximizers and they make two decisions: i) which service to choose, and ii) if BE is chosen, how much to bid.

Let i denote the service class such that $i=1$ for G service and $i=2$ for BE service. G service is offered with a fixed price p_G per unit time and each customer paying this price gets a dedicated resource. The price for BE service, $p_{BE}(t)$, is a RCLL (right-continuous with left limits) discrete-level stochastic process in the interval $[\underline{p},\overline{p}]$ with N price levels ($\underline{p}=p_N\leq p_{N-1}\leq\ldots\leq p_1=\overline{p}$). We will not characterize the dynamics at this stage, but assume that users with infinite service level requirements decide based on the steady state probability mass function associated with $\{p_{BE}(t),t\geq 0\}$ which is denoted by $\boldsymbol{\pi}=(\pi_1,\pi_2,\ldots,\pi_N)$, and assumed to exist, has support $\mathcal{P}=[\underline{p},\overline{p}]\in(0,\infty)$. In this option, customers place their bids and the SP offers service to each BE user whose bid is larger than or equal to the prevailing spot price $p_{BE}(t)$, and interrupts service to all bidders below $p_{BE}(t)$. That is, a BE user that bids b is *active* $\forall t$ s.t. $p_{BE}(t)\leq b$ and *interrupted* $\forall t$ s.t. $p_{BE}(t)>b$. We assume interruptions have no cost to the SP and an interrupted job resumes without any additional setup cost (if service disruptions are infrequent and service times are long—infinite in our model—this modeling idealization may be reasonable). It is worth noting that in our infinite capacity model the price dynamics are controlled by the SP as opposed to stochastic supply-demand imbalance effects.

Users are heterogeneous and characterized by their idiosyncratic (monetary) valuation per unit time of receiving service v and disutility (congestion sensitivity) parameter κ, which measures the monetary loss per unit of time where the service is unavailable. Consider a user with valuation v and congestion sensitivity parameter κ and bid \$$b$ for BE service. For a user that selects BE service and bids \$$b$, we will define $\alpha(b)$ to be the fraction of time her service is active, and $p(b)$ to be the payment per unit time:

$$\alpha(b)=\sum_{i:p_i\leq b}\pi_i \qquad \text{and} \qquad p(b)=\sum_{i:p_i\leq b}\pi_i p_i.$$

The net utilities for the two service options are:

$$U_1(v,\kappa)=v-p_G \qquad \text{and} \qquad U_2(v,\kappa,b)=\alpha(b)v-\kappa(1-\alpha(b))-p(b).$$

That is, customers extract the per unit value \$$v$ when their service is active; when their service is interrupted they forgo this value and incur a cost of \$$\kappa$ per unit time. They only pay while their service is active, captured by p_G and $p(b)$ for each option, respectively.

The optimal BE bid for a user with parameters v and κ is

$$b(v,\kappa) = \arg\max_{\underline{p}\le b\le \overline{p}} U_2(v,\kappa,b).$$

LEMMA 1. *Without loss of generality,* $b(v,\kappa) \in \{p_1, p_2, \ldots, p_N\}$.

Moreover, if there are multiple bids that achieve the maximum, we assume that the lowest maximizing bid is selected. We let $U_2(v,\kappa) := U_2(v,\kappa,b(v,\kappa))$. A user with parameters v and κ chooses service-$i^*(v,\kappa)$, where

$$i^*(v,\kappa) = \arg\max_{i=1,2} \{U_i(v,\kappa) : U_i(v,\kappa) \ge 0\} \text{ and set } i^*(v,\kappa) = 0 \text{ if } U_i(v,\kappa) < 0 \text{ for } i = 1,2,$$

where $i = 0$ represents the no-buy option.

As mentioned in the introduction the monotonicity of v/κ as κ grows plays a crucial role. In the following two sections, we will study this infinite capacity stylized model, and, specifically consider the SP's revenue maximization problem, and we will consider two settings: first sub-linear then super-linear increase in valuation with congestion sensitivity. We will consider two models of customer heterogeneity (v,κ). In Section 4.1.1–4.1.2, we will consider that types are continuous; in Sections 4.2 and 5, we will consider a discrete model.

4. Sub-Linear Increase in Valuation with Congestion Sensitivity

We study a market where the valuation rate v grows more slowly than her corresponding congestion sensitivity parameter κ; i.e., users with increasing congestion sensitivity may indeed value the service more, but their valuation does not grow as fast as the corresponding disutility from service interruption.

4.1. Linear Dependence between valuation and congestion rate (v,κ)

We consider a continuum of user types indexed by η. A type η user has a positive willingness-to-pay $v := A + \eta$ per unit time of service and a positive congestion sensitivity parameter $\kappa := B\eta$, where A, B are positive constants common across all consumers. User types are assumed to be independent and identically distributed (i.i.d.) draws from a continuous distribution F with density f, which is assumed strictly positive and continuously differentiable on the interval $\mathcal{N} = [\underline{\eta}, \overline{\eta}] \subseteq [0,\infty)$. Let $\bar{F} = 1 - F$. Hence, v and κ are linearly dependent and user heterogeneity is one dimensional. Note that in this setting, both the

valuation rate ($v = A + \eta$) and the congestion rate ($\kappa = B\eta$) are increasing function of the user type η, and that relative rate of growth of $v/\kappa = \frac{A+\eta}{B\eta}$ is decreasing in their type. We summarize this model for ease of reference below:

$$\text{Model 1: } v = A + \eta,\ \kappa = B\eta,\ A, B > 0,\ \eta \sim F. \tag{1}$$

4.1.1. BE randomizes between 2 price levels (high/low). The SP will offer the BE service at two price levels $\$p_H$, $\$p_L$ with $p_H \geq p_L$, and choose π, the fraction of time the BE service is priced at $\$p_L$. A customer that bids $\$p_L$ for BE will enjoy the service for π fraction of time, and if she bids $\$p_H$, she enjoys the service without interruption, and pays $\pi p_L + (1-\pi)p_H$. From Lemma 1, customers do not bid any other amount. Guaranteed service is priced at $\$p_G$. Without loss of generality we will assume that $\pi p_L + (1-\pi)p_H > p_G$, that is, if a user wants guaranteed service, then she will choose the G service option at $\$p_G$. We will add this as a constraint to our downstream revenue optimization formulation.

The utilities for two services can be written as a function of η as

$$U_1(\eta) = (A+\eta) - p_G \text{ and } U_2(\eta) = \pi(A+\eta) - B\eta(1-\pi) - \pi p_L = \pi(A+\eta-p_L) - B\eta(1-\pi).$$

$U_i(\eta)$ takes the form of $U_i(v,\kappa)$ in this model with one dimensional user types.

We will first assume that the utility gained from the BE service is non-decreasing in η for any η, i.e.,

$$U_2'(\eta) \geq 0 \iff \pi - B + B\pi \geq 0 \iff \pi \geq \frac{B}{1+B}, \tag{2}$$

which implies a constraint on the choice of π to the SP.

Later on we will formulate and solve the problem for the case $\pi < \frac{B}{1+B}$ and show that the respective solution is sub-optimal. Let

$$\mathcal{S}_G = \{\eta | U_1(\eta) \geq U_2(\eta), \text{ and } U_1(\eta) \geq 0\} \text{ and } \mathcal{S}_{BE} = \{\eta | U_2(\eta) > U_1(\eta), \text{ and } U_2(\eta) \geq 0\},$$

denote the sets of customer types that choose G and BE service, respectively. From (2) and the fact that $U_1'(\eta) \geq U_2'(\eta)$ for any $B > 0$ and $0 \leq \pi \leq 1$, we get that

$$\mathcal{S}_G = \{\eta | \eta \geq \eta_H \text{ and } \eta \geq p_G - A\} \text{ and } \mathcal{S}_{BE} = \{\eta | \eta < \eta_H \text{ and } \eta \geq \eta_L\},$$

where η_H and η_L satisfy

$$(1+B)(1-\pi)\eta_H = p_G - \pi p_L - (1-\pi)A \text{ and } (\pi - B(1-\pi))\eta_L = \pi(p_L - A). \tag{3}$$

That is, customer type η chooses G if $\eta \geq \eta_H$ and $\eta \geq p_G - A$, chooses BE if $\eta_L \leq \eta < \eta_H$, and does not join the system if $\eta < \eta_L$. The marginal types η_L, η_H are controlled by the SP through p_G, p_L, p_H, and π. Here we are restricting our analysis to the case that $\eta_H \geq \eta_L$. If $\eta_L > \eta_H$, then the BE service becomes unattractive, and the SP is offering only G service (this is also the case when $\eta_L = \eta_H$); Based on this observation, we can disregard from consideration the case where $\eta_L > \eta_H$.

We will first assume that $\eta_H \geq p_G - A$ and formulate and solve SP's revenue maximization problem. Then we will show that any solution with $\eta_H < p_G - A$ is sub-optimal and verify the assumption is satisfied under the optimal solution.

Assuming $\eta_H \geq p_G - A$, the revenue function of the SP is

$$\begin{aligned} R_1 &= p_G \bar{F}(\eta_H) + \pi p_L (F(\eta_H) - F(\eta_L)) \\ &= (p_G - \pi p_L)\bar{F}(\eta_H) + \pi p_L \bar{F}(\eta_L) \\ &= \left[\eta_H(1+B)(1-\pi) + (1-\pi)A\right]\bar{F}(\eta_H) + \left[\pi(A+\eta_L) - B\eta_L(1-\pi)\right]\bar{F}(\eta_L) := R(\eta_H, \eta_L, \pi). \end{aligned}$$

The SP's revenue maximization problem is:

$$\underset{\eta_H, \eta_L, \pi}{\text{maximize}} \quad R(\eta_H, \eta_L, \pi) \tag{4}$$

$$\text{subject to} \quad \eta_L \leq \eta_H,\ \pi \geq \frac{B}{1+B},\ \pi \leq 1. \tag{5}$$

In contrast, if $\eta_H < p_G - A$, the revenue function reduces to

$$R_2 = p_G \bar{F}(p_G - A) + \pi p_L (F(\eta_H) - F(\eta_L)) \leq p_G \bar{F}(\eta_H) + \pi p_L (F(\eta_H) - F(\eta_L)) = R_1,$$

and the corresponding constraint set is smaller than in (5). It follows that any solution with $\eta_H < p_G - A$ is sub-optimal.

Next we solve (4)–(5) in terms of η_H, η_L, and π. These three parameters uniquely determine p_G and p_L from (3), and we show that the optimal solution satisfies $p_G \geq p_L$.

PROPOSITION 1. *Consider the model specified by (1) and let $(\eta_H^*, \eta_L^*, \pi^*)$ be an optimal solution to (4)–(5), and p_G^* and p_L^* be the optimal prices corresponding to the solution triple. Then,*

1. $\pi^* = \dfrac{B}{1+B}$,
2. $\eta_L^* = \underline{\eta}$ *with* $p_L^* = A$,

3. $\eta_H^* = p_G^* - A$.

(All proofs are given in the Appendix.) We can simplify the revenue maximization problem to

$$\underset{\eta_H}{\text{maximize}} \quad \left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H) + \frac{B}{1+B}A. \tag{6}$$

PROPOSITION 2. *Under the model specified by (1), it is optimal to offer G and BE services if and only if*

$$f(\underline{\eta}) < \left(\underline{\eta} + \frac{A}{1+B}\right)^{-1}. \tag{7}$$

Once η_H^* and π^* are identified, the optimal price pair (p_G^*, p_L^*) can be chosen so as to satisfy Proposition 1. We mentioned earlier that without loss of generality we will restrict attention to prices such that

$$(1-\pi^*)p_H^* + \pi^* p_L^* > p_G^*, \tag{8}$$

i.e., it is sub-optimal for users that want uninterrupted service to choose BE but submit a high bid ($\$p_H$). Any choice of p_H that satisfies (8) will suffice.

To establish that Proposition 1 indeed characterizes the globally optimal solution, we need to rule out any solution where $\pi < \dfrac{B}{1+B}$ and, as a consequence, $U_2(\eta)$ is decreasing in η. If $U_2(\underline{\eta}) \le 0$, there is no BE service, i.e., reducing to a one-service solution. Assuming $U_2(\underline{\eta}) > 0$, $\mathcal{S}_G$ and $\mathcal{S}_{BE}$ can be written as

$$\mathcal{S}_G = \{\eta | \eta \ge \eta_H \text{ and } \eta \ge p_G - A\} \text{ and } \mathcal{S}_{BE} = \{\eta | \eta < \eta_H \text{ and } \eta \le \eta_L\},$$

where η_H and η_L satisfy (3). Then the SP's revenue maximization problem is:

$$\underset{\eta_H,\eta_L,p_G,\pi}{\text{maximize}} \quad p_G \cdot \int_{\eta\in\mathcal{S}_G} f(\eta)d\eta + \pi p_L \cdot \int_{\eta\in\mathcal{S}_{BE}} f(\eta)d\eta \quad \text{subject to} \quad 0 \le \pi < \frac{B}{1+B}. \tag{9}$$

PROPOSITION 3. *Consider the model specified by (1). The optimized revenue rate for (9) is bounded above by the optimized objective in (6). Therefore, $\pi < \dfrac{B}{1+B}$ is sub-optimal.*

4.1.2. Can the SP do better by offering BE with $N>2$ price levels? Consider now the case where the SP will offer the BE service on an N-price grid given by $p_1 \geq p_2 \ldots, \geq p_N \geq 0$ and let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$, where π_i is the fraction of time the prevailing BE service price is p_i. As in the two-price-level analysis, p_1 can be set at a sufficiently high value to ensure that users that would bid p_1 (or higher) to enjoy uninterrupted BE service, would prefer instead to pay p_G and get the guaranteed service option. Using Lemma 1, we only consider the bids that are on the price grid $\{p_1, p_2, \ldots, p_N\}$. Define $\bar{\pi}_k = \sum_{j=k}^{N} \pi_j$ and $\bar{p}_k = \sum_{j=k}^{N} \pi_j p_j$, and redefine the net utility function for BE as

$$U_2(\eta, p_k) = \bar{\pi}_k(A+\eta) - B\eta(1-\bar{\pi}_k) - \bar{p}_k, \quad k = 2, \ldots, N$$

for a type η customer bidding p_k.

As in the two-price-level case, we assume that the utility gained from the BE service is non-decreasing in η for any η and $\bar{\pi}_i$ values, i.e.,

$$\begin{aligned} U_2'(\eta, p_i) \geq 0,\ i = 2, 3, \ldots, N &\iff \bar{\pi}_i \geq \frac{B}{1+B},\ i = 2, 3, \ldots, N \\ &\iff \pi_N \geq \frac{B}{1+B}. \end{aligned}$$

Let $\mathcal{S}_G$ denote the interval for customer types that choose G service, and sets $\mathcal{S}_{BE}^i$ the interval for customer types that choose BE service and bid p_i $(i = 2, 3, \ldots, N)$:

$$\begin{aligned} \mathcal{S}_G &= \{\eta | U_1(\eta) \geq U_2(\eta, p_i),\ i = 2, 3, \ldots, N \text{ and } U_1(\eta) \geq 0\} \text{ and} \\ \mathcal{S}_{BE}^i &= \{\eta | U_2(\eta, p_i) \geq U_2(\eta, p_k),\ k \neq i; U_2(\eta, p_i) \geq U_1(\eta), \text{ and } U_2(\eta, p_i) \geq 0\},\ i = 2, 3, \ldots, N. \end{aligned}$$

Then the SP's revenue maximization problem is:

$$\underset{p_G, p_2, p_3, \ldots, p_N, \boldsymbol{\pi}}{\text{maximize}} \quad p_G \cdot \int_{\eta \in \mathcal{S}_G} f(\eta) d\eta + \sum_{i=2}^{N} \bar{p}_i \cdot \int_{\eta \in \mathcal{S}_{BE}^i} f(\eta) d\eta \tag{10}$$

$$\text{subject to} \quad \pi_N \geq \frac{B}{1+B},\ 1^T \boldsymbol{\pi} = 1,\ \boldsymbol{\pi} \geq 0. \tag{11}$$

PROPOSITION 4. *Consider the model specified by (1) and let k^* be the number of distinct price levels offered in BE service at the optimal solution of (10)–(11). Then, an optimal solution is to use $k^* = 2$ with the structure specified in Proposition 1.*

Once again, for the model in (1) with the affine relation between (v, κ), it is optimal to offer G service and BE service with two-price-level if and only if (7) is satisfied.

4.2. General Dependence Between Valuation and Congestion Sensitivity

So far we have restricted attention to the affine dependence between (v, κ) that allowed us to solve the resulting revenue maximization problem in closed form. In this subsection we briefly consider a market where the (v, κ) dependence is general, yet still v/κ grows sub-linearly with respect to κ, and primarily show that in such a setting the SP may wish to offer more than 2 price levels for BE service.

Suppose there are n customer types and N price levels in BE service ($N > n$). Let $\kappa_1 > \kappa_2 > \ldots > \kappa_n > 0$ with $v_1 \geq v_2 \geq \ldots \geq v_n > 0$ such that $\frac{v_1}{\kappa_1} < \frac{v_2}{\kappa_2} < \ldots < \frac{v_n}{\kappa_n}$. The fraction of users that are of type i is λ_i. Let $p_1 \geq p_2 \geq \ldots \geq p_N \geq 0$ be the price levels with $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ such that π_j is the fraction of time the system is in price level p_j ($\mathbf{1}^T\boldsymbol{\pi} = 1$ and $\boldsymbol{\pi} \geq 0$).

$$\text{Model 2: } v_1 \geq v_2 \geq \ldots \geq v_n > 0,\ \kappa_1 > \kappa_2 > \ldots > \kappa_n > 0,\ \frac{v_1}{\kappa_1} < \frac{v_2}{\kappa_2} < \ldots < \frac{v_n}{\kappa_n}. \tag{12}$$

The objective of the SP is to maximize its revenue rate by offering a price vector $(p_G, \boldsymbol{p})$ and an availability vector $\boldsymbol{\pi}$. As previously, one can restrict attention to customer bids in $\{p_1, p_2, \ldots, p_N\}$.

PROPOSITION 5. *Consider the model specified by (12) and let p^* be the optimal price when there is only G service. Then, it is optimal to offer G and BE services together if and only if $p^* > v_n$.*

Proposition 5 shows that if some customer types choose not to buy under the optimal single service level (only G) solution, then the SP can extract more revenue by offering the second service level (BE). We have shown above that when there is linear dependence between v and κ, it is enough to offer BE service with two price levels. The example given below shows that when (v, κ) have a general dependence structure and still v grows sub-linearly with respect to κ, i.e., the model in (12), it may be optimal to use $k > 2$ price levels. In particular, the following example shows that the property proven in Proposition 4 no longer holds (the relation between (v, κ) is quadratic):

Example: Three customer types with $\boldsymbol{\lambda} = (1, 1, 1)$, $\boldsymbol{v} = (4, 2, 1)$, $\boldsymbol{\kappa} = (16, 4, 1)$. The optimal solution is $p_G = 4$, $\boldsymbol{p} = (p_1, 6, 2/3)$, $\boldsymbol{\pi} = (1/7, 3/28, 3/4)$ where $p_1 > 20$.

5. Super-Linear Increase in Valuation with Congestion Sensitivity

This section uses the discrete type (v,κ) model of Section 4.2 but assumes that $\frac{v_1}{\kappa_1} \geq \frac{v_2}{\kappa_2} \geq \ldots \geq \frac{v_n}{\kappa_n}$, i.e., that the valuation rate grows super-linearly with respect to the corresponding congestion sensitivity rate.

$$\text{Model 3: } v_1 \geq v_2 \geq \ldots \geq v_n > 0,\ \kappa_1 > \kappa_2 > \ldots > \kappa_n > 0,\ \frac{v_1}{\kappa_1} \geq \frac{v_2}{\kappa_2} \geq \ldots \geq \frac{v_n}{\kappa_n}. \tag{13}$$

The objective of the SP is to maximize its revenue by offering price vector $\boldsymbol{p}$ and availability vector $\boldsymbol{\pi}$. Similar to Lemma 1, in this setting users need only consider bids that are equal to one of the offered price points. Here we do not introduce a separate G service at first. User types that bid equal to the highest price level, p_1, receive uninterrupted service (i.e., G service) and pay $\overline{p}_1$. Hence, we first consider only BE service first, find the optimal pricing mechanism in this case, and then introduce a separate G service with price $\overline{p}_1$. The next proposition characterizes the structure of the optimal solution when the SP maximizes its revenue over the price grid and associated π's.

PROPOSITION 6. *Consider the model specified by (13). Let k^* be the number of distinct price levels offered in BE service. Then, $k^* \leq 2$.*

Proposition 6 shows that it is optimal to offer BE service with at most two price levels. Let these price levels be p_H and p_L with $p_H \geq p_L$, and π is the fraction of time BE service is priced at p_L. If $p_H = p_L$, then the solution has only one service level, which is uninterrupted service. If $p_H > p_L$, then customers that bid p_H enjoys uninterrupted service by paying $\pi p_L + (1-\pi)p_H$. In this case, in addition to BE service we can offer G service with price $p_G := \pi p_L + (1-\pi)p_H$. Customer types bidding p_H previously are now indifferent between bidding p_H for BE service or paying p_G for G service. To ensure that these customer types choose G service over BE service with bid p_H, we can increase p_H without changing p_G. Now, these customers are no longer indifferent between the two options. Note that this change would not affect the choice of customer types that choose to bid p_L.

Next, we compare the optimal revenue that the SP makes with at most two service levels, i.e., two price levels, with the optimal revenue under one service level. Proposition 7 shows that the revenue under the former case is bounded above by the latter, or equivalently, offering one price level is optimal.

PROPOSITION 7. *Consider the model specified by (13). Then, $k^* = 1$.*

Offering BE service with one price level means that the price is constant over time and customer types that bid this constant price get uninterrupted service, which is equivalent to G service, and the rest of the customer types do not get any service. Thus, we conclude that offering only G service is optimal if (v, κ) follows (13).

The result above is consistent with the policy identified in Katta and Sethuraman (2005), whereat the authors showed that for the model considered in this section that optimal policy for a SP operating a system with congestion effects (arising through the operation of an $M/M/1$ system) is optimal not to inject any strategic delay. In a large scale system, the service level that arises due to stochastic congestion effects becomes small, and in an infinite capacity system altogether disappears.; i.e., if the SP can avoid congestion effects, she will indeed select to do so. This is what we see in our model as well. (Similarly to what we mentioned earlier the structure of the user utility function is different in our model, so a direct application of these earlier findings is not possible.)

6. Data

We first offer a description of price data from AWS (a cloud computing platform offered by Amazon), and then offer a brief calibration and discussion of our model on AWS data.

6.1. Descriptive Statistics

Amazon is the biggest cloud computing SP. They offer over 1,000 products to the IaaS market in 9 regions globally. For each product, the price trace of the last 90 days is made publicly available by Amazon. We have obtained price traces from August 2013 onwards for the spot instances using an automated script that we programmed, which runs everyday and downloads and stores the price traces of the last 24 hours, for all products. This script has enabled us to have a longer time frame for the price history. Amazon does not disclose any information other than the price traces.

We have analyzed the data traces from March 1, 2015 to August 31, 2015 for 1,122 products. The products are categorized under five different classes by AWS: "compute optimized," "general purpose," "GPU instances," "memory optimized," and "storage optimized." In each of these classes there are multiple machine sizes. Moreover, prices differ with respect to the location of the product and the operating system the product has. To facilitate reporting statistics on pools of different products with different on-demand prices, we normalize the spot and on-demand prices of each product by the respective on-demand

price. In this manner, a normalized spot price is unit-less and expressed and understood as a multiple of the underlying on-demand price; all products have a normalized on-demand price equal to 1. To get a better sense of pricing dynamics, first we look at the descriptive statistics per product. For each product, we calculate: the average normalized spot price; the normalized spot price range; the average uptick and downtick inter-arrival times; the average magnitudes of the corresponding spot price jumps; and, the fraction of time the spot price is greater than on-demand price. Table 2 shows that the mean of the average normalized spot prices across products is about half of the on-demand price. More than 92% of the products have a time-average spot price less than 1, which means that for more than 92% of the products, procuring spot instances, with sufficiently high bids so as never to be shut off, would cost less than on-demand instances for the whole 6-month period. We discuss this result more in Section 6.2. Further, summary statistics shows that the range of spot price fluctuations is wide, more than three times of the corresponding on-demand price on average. The average inter-arrival time of an uptick (downtick) price change is in the order of hours, and the average magnitude of an uptick (downtick) is about one third of the on-demand prices. Lastly, for most products the spot price is below the corresponding on-demand price for more than 90% of the time. Figure 1 shows the distribution of each of these categories (with a few outliers discarded in each plot).

Table 2 Summary of descriptive statistics per product

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Average normalized spot price	0.037	0.213	0.366	0.515	0.780	3.756
Normalized spot price range	0.000	0.829	1.511	3.216	4.772	39.050
Avg. uptick inter-arrival time (hrs)	0.000	3.071	7.115	35.200	27.930	1428.000
Avg. downtick inter-arrival time (hrs)	0.000	3.088	6.924	33.950	27.200	1111.000
Average uptick magnitude	0.000	0.144	0.284	0.449	0.537	9.740
Average downtick magnitude	0.000	0.143	0.290	0.459	0.537	12.150
Fraction of time spot>on-demand	0.000	0.000	0.008	0.072	0.066	1.000

Next, we assume that a user selects spot service and bids sufficiently high so that she is never outbid and would enjoy uninterrupted service. For each different possible time of arrival, we record the average price she would pay per hour if she stayed in the system for

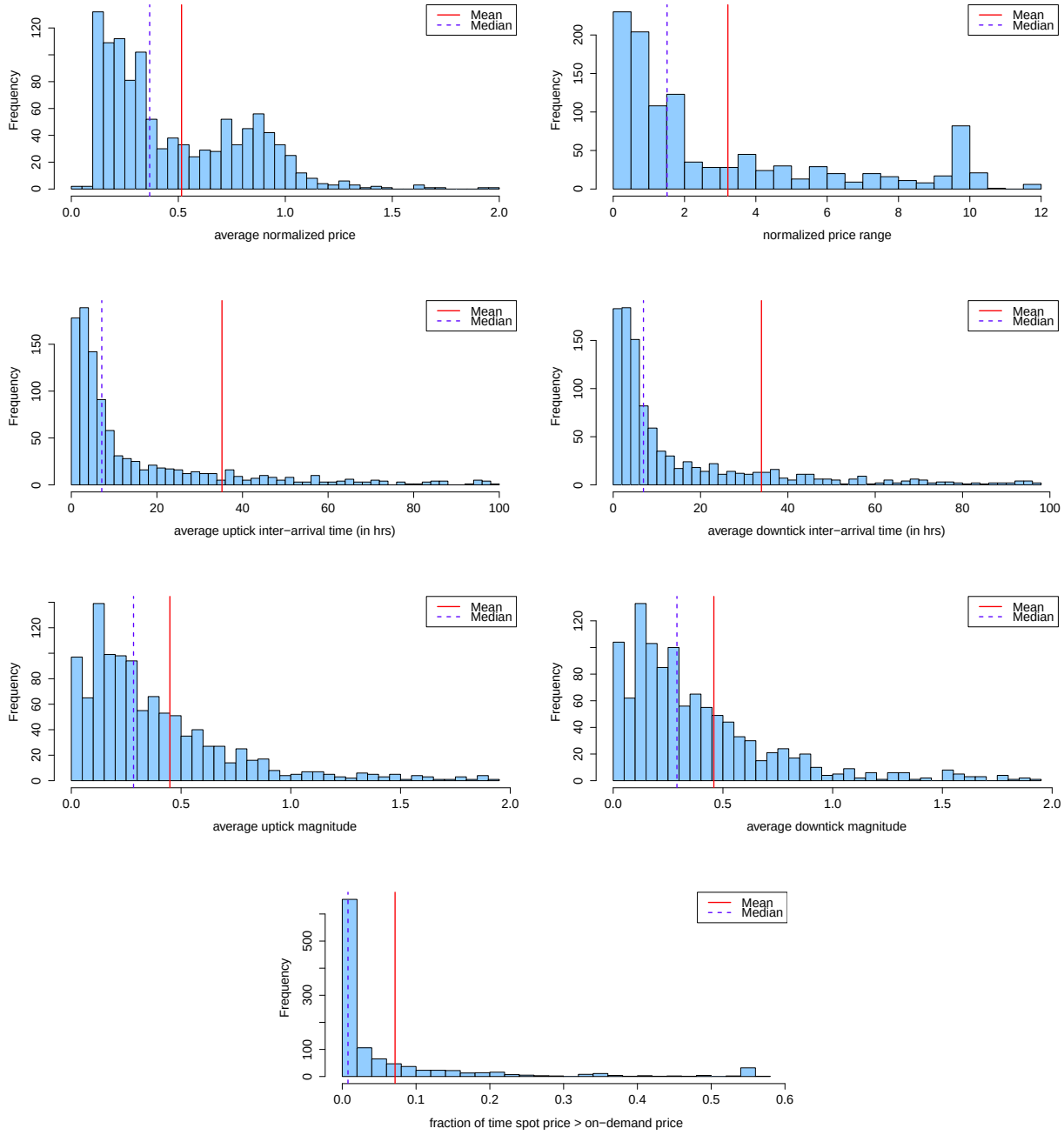

Figure 1 **Histogram of descriptive statistics per product**

1 hour, 1 day, 1 week, or 1 month. For each of the 4 usage durations, we average across time of arrival. The results are reported in Figure 2. These plots show that most Windows products have higher spot prices compared to Linux/UNIX products, i.e., the potential gain from the spot market is less for Windows products. The four panels in Figure 2 are

similar, suggesting that the usage duration does not play an important role on the selection between spot (at maximum bid) versus on-demand.

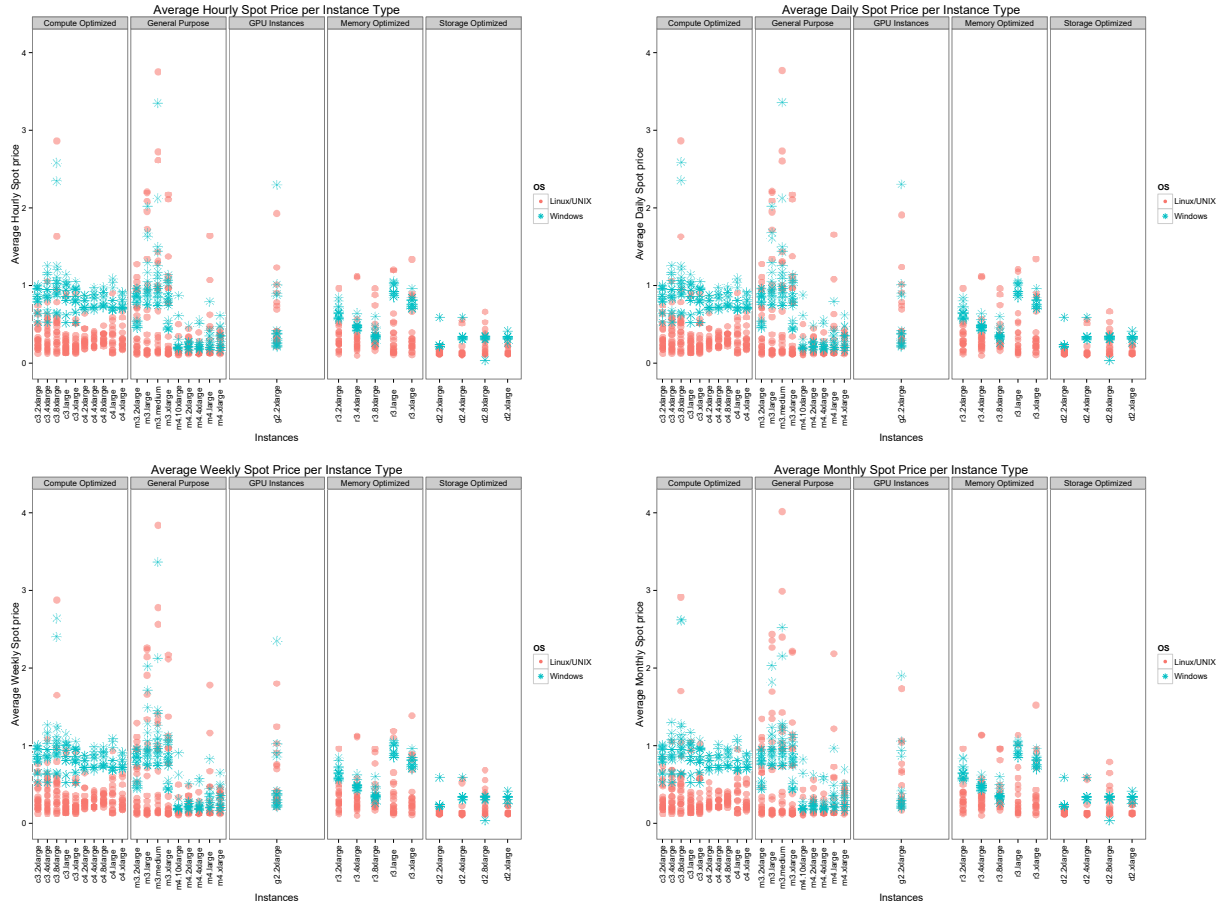

Figure 2 Average prices for different duration of usage

To illustrate the fluctuations in the prevailing spot prices over time, we focused on the running averages of the prevailing spot prices for daily, weekly, and monthly usage updated daily for every class of product. Figure 3 summarizes how the average spot price fluctuates over time under daily, weekly, and monthly usage of "GPU Instances" for Linux/UNIX and Windows machines; there are 18 such products for each operating system in total. While the solid line represents the average price of all products in this class, the blue (shaded) area denotes $\pm$ one standard deviation band from the average price (computed across the respective 18 data points in each time point). For this product class, prices for both operating systems follow similar patterns whereat the spot market is cheaper than the on-demand market until the beginning of August. The standard deviation also increased

during that period, implying also increased variation across different "GPU Instances" products during this peak period. We observe different patterns in other product classes.

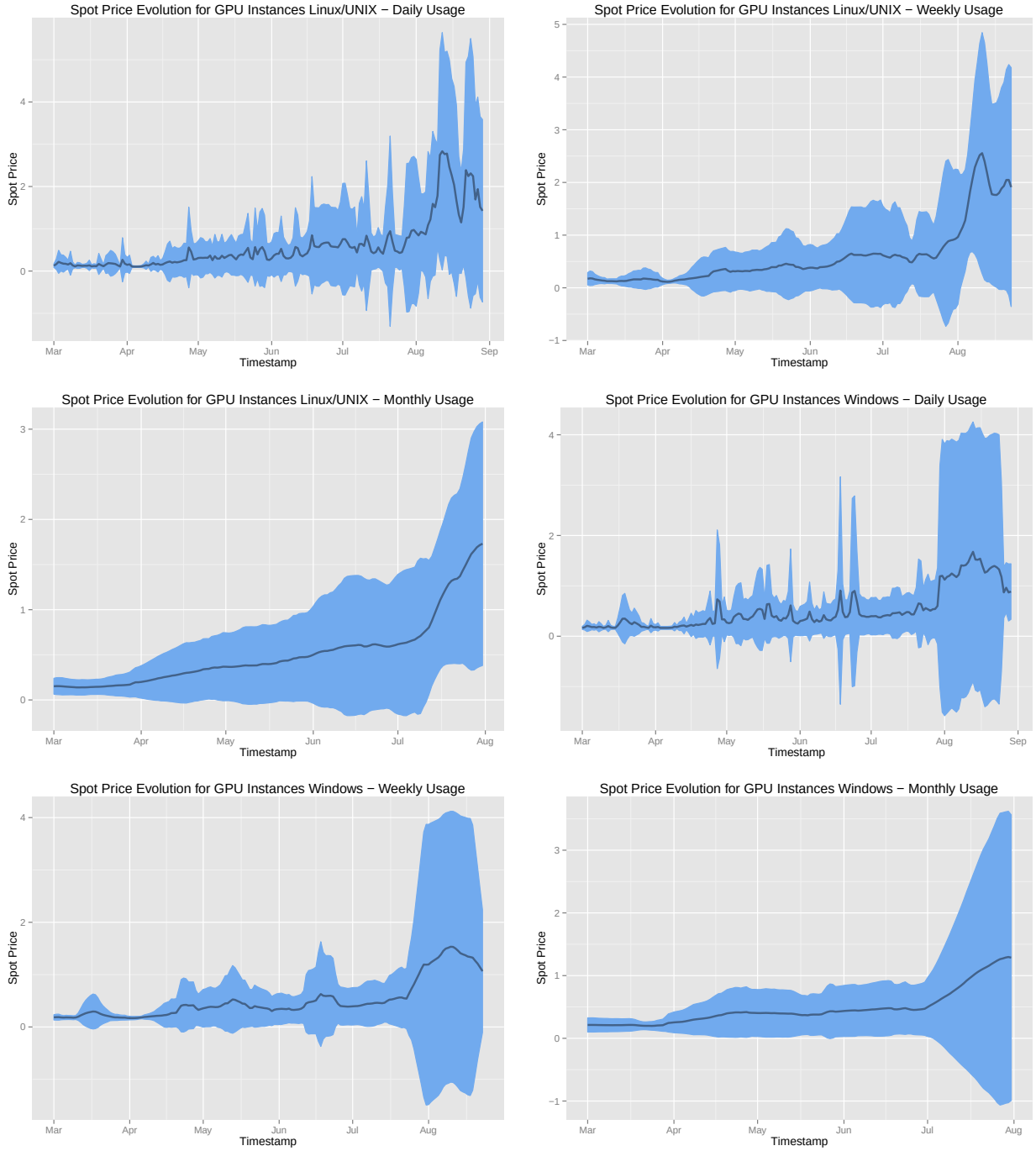

Figure 3 **Average price change over time with different usage times**

6.2. Data Evidence

We can use the AWS data to calibrate our model primitives. To repeat, the model of Section 4.1 assumes a linear dependence between valuation and congestion rates, which implies that valuation grows sub-linearly relative to the congestion rate. The results of Section 4.1 suggest that it suffices for the SP to offer the spot price service with just two price levels (high/low). We proceed as follows. We assume the model specified in Section 4.1 is in force and that the SP follows the optimal policy derived in Section 4.1. We compute the empirical spot price occupancy distribution, and approximate it with a two-level (p_H, p_L, π) distribution. We then derive the implied user valuation and congestion model parameters. We will approximate the empirical distribution by the triple (p_H, p_L, π) that is closest in the sense of the Kantorovich metric, where p_H is the high price, p_L is the low price, and π is the fraction of time the price is equal to p_L. (The Kantorovich metric between two random variables X and Y in $\mathbb{R}$ is defined as $K(X,Y) = \int_{\mathbb{R}} |F_X(x) - F_Y(x)|\, dx$, where F_X and F_Y are the cumulative distribution function of X and Y, respectively.)

Assuming that user types are distributed uniformly on $(0, \eta_{max})$, the parameters $A, B, \eta_{\max}$ can be calculated using the results from Section 4.1. Specifically, for given (p_H, p_L, π),

$$B = \frac{\pi}{1-\pi},\ A = p_L,\ \eta_{max} = 2(p_G - A) + \frac{A}{1+B}$$

where $p_G = 1$ since the price path is normalized based on the G service price. Using these parameters we can get the implied valuation and congestion sensitivity parameter for each user type η: her valuation rate is equal to $A + \eta$ and her congestion sensitivity parameter is equal to $B\eta$. Lastly, we analyze if the implied parameters satisfy the conditions on p_H and p_L, i.e. whether $p_G < (1-\pi)p_H + \pi p_L$, $p_H > 1$, $p_L < 1$ hold. Our analysis containing the data for the period Mar. 1, 2015 – Aug. 31, 2015 has shown that out of 1,122 products, only 63 of them satisfy all these conditions. The summary statistics of the normalized price path of these 63 products is given in Table 3. Calibrating our model on the observed data we get the parameters shown in Table 4.

The normalized estimated parameters suggest the following:

- Valuation per unit time: $(A + \eta) \sim U(0.6, 1.5)$.
- Congestion cost per unit of downtime: $B\eta \sim U(0, 12.5)$. Specifically, we note that congestion costs when the system is down, due to lost revenue and possibly lost goodwill/reputation, can be up to 4x-10x of the valuation per unit time.

Table 3 Summary statistics of price paths

Avg. Price	Price Range	Reserve Price	Price>1
1.446	6.438	0.608	29%

Table 4 Estimated parameters on average

A	B	η_{max}	η_H	p_L	p_H	π
0.638	14.992	0.835	0.362	0.638	5.291	0.784

- Fraction of downtime: $1-\pi=0.216$.
- Congestion cost per unit time (due to service interruption): $\sim U(0,2.7)$.
- Lowest valuation per unit time choosing G: $A+\eta_H=1$.

These parameter estimates suggest that for high customer types, the disutility from service disruption in spot service is of the same order of magnitude (or higher) as the valuation itself, and as a result, only the least congestion-sensitive users will choose that option. In our data, this seems to be the lower 40% of the distribution that wants G service.

Finally, as noted earlier, for more than 92% of the products, a user would be better off selecting the spot option and bid sufficiently high so as to receive continuous uninterrupted service for the whole 6-month period. Based on our model, this would suggest insufficient degradation of the spot service option by the SP so as to incentivize congestion sensitive customers to choose the on-demand service option. Assuming the estimated parameters on Table 4 also hold for all offered products and the demand for each of these products is the same, our back-of-the-envelope calculation shows that Amazon could almost double the revenue extracted from these products by further optimizing the pricing of the spot option. Of course, this calculation disregards other (unmodeled) economical and technological considerations that may affect such tactical pricing decisions, and for which we lack transparent data.

Acknowledgments

This research was supported by a grant from the W. Edwards Deming Center at Columbia Business School.

Appendix. Proofs

Proof of Proposition 1 1. Suppose $\pi^*=\dfrac{B}{1+B}+\varepsilon$, $\varepsilon>0$ and $\bar{\pi}=\dfrac{B}{1+B}$, where $\bar{\pi}$ is not one of the optimal values for π. Then,

$$R(\eta_H^*,\eta_L^*,\bar{\pi})-R(\eta_H^*,\eta_L^*,\pi^*)=\varepsilon\left\{\left[(1+B)\eta_H^*+A\right]\bar{F}(\eta_H^*)-\left[(1+B)\eta_L^*+A\right]\bar{F}(\eta_L^*)\right\}<0$$

$$\iff [(1+B)\eta_H^* + A]\,\bar{F}(\eta_H^*) < [(1+B)\eta_L^* + A]\,\bar{F}(\eta_L^*)$$
$$\implies \eta_H^* = \eta_L^* + \gamma\ (\gamma > 0)$$

However, decreasing η_H^* by γ increases R. Therefore, $(\eta_H^*, \eta_L^*, \pi^*)$ is not the optimal solution. Contradiction.

2. If $\pi^* = \frac{B}{1+B}$, then the willingness to pay for BE service is $A\frac{B}{1+B}$, which is independent of the customer types. Therefore, η_L^* is either $\underline{\eta}$ or $\overline{\eta}$. If $\eta_L^* = \overline{\eta}$, then $\eta_H^* \geq \overline{\eta}$ since $\eta_L^* \leq \eta_H^*$. However, $R(\overline{\eta}, \overline{\eta}, \frac{B}{1+B}) = 0$, and $R(\eta_H, \eta_L, \frac{B}{1+B})$ is nonnegative for all $\underline{\eta} \leq \eta_L \leq \eta_H \leq \overline{\eta}$. Therefore, $\eta_L^* = \underline{\eta}$. Finally, if $\eta_L^* = \underline{\eta}$, then $p_L^* = A$.

3. $\eta_H^* = \eta_H^*(1+B)(1-\pi^*) = p_G^* - A + \pi^*(A - p_L^*) \iff \eta_H^* - \pi^*(A - p_L^*) = \eta_H^* = p_G^* - A$.

Proof of Proposition 2 The first order condition for the objective function is

$$\frac{d}{d\eta_H}\left[\left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H)\right]\Big|_{\eta_H=\eta_H^*} = 0$$
$$\Rightarrow \bar{F}(\eta_H^*) - \left(\eta_H^* + \frac{A}{1+B}\right) f(\eta_H^*) = 0$$
$$\Rightarrow f(\eta_H^*) = \frac{\bar{F}(\eta_H^*)}{\eta_H^* + \frac{A}{1+B}}$$

$\Rightarrow$: If there are two services offered, then $\eta_H > \underline{\eta}$, which implies

$$\frac{d}{d\eta_H}\left[\left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H)\right]\Big|_{\eta_H=\underline{\eta}} > 0 \quad \Rightarrow f(\underline{\eta}) < \frac{1}{\underline{\eta} + \frac{A}{1+B}}$$

$\Leftarrow$: Suppose (7) holds and the SP offers only G. This implies $\eta_H^* = \underline{\eta}$. However, if (7) holds

$$\frac{d}{d\eta_H}\left[\left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H)\right]\Big|_{\eta_H=\eta_H^*} > 0,$$

which contradicts the optimality condition of η_H.

Proof of Proposition 3 We allow $0 \leq \pi \leq \frac{B}{1+B}$ instead of strict inequality and show the proposition holds for this larger feasible region. We analyze four collectively exhaustive alternatives for the bounds of S_G and S_{BE} below.

- *Case 1: $\eta_L \geq \eta_H$ and $\eta_H \geq p_G - A$:* The revenue function becomes

$$\begin{aligned}\bar{R}_1 &= [\eta_H(1+B)(1-\pi) + \pi\eta_L - B(1-\pi)\eta_L + A]\,\bar{F}(\eta_H) + [\pi\eta_L - B(1-\pi)\eta_L + A\pi]\,F(\eta_H) \\ &= [\eta_H + (\eta_L - \eta_H)(\pi - B(1-\pi)) + A]\,\bar{F}(\eta_H) + [\pi\eta_L - B(1-\pi)\eta_L + A\pi]\,F(\eta_H)\end{aligned}$$

$\bar{R}_1$ is non-decreasing in π. Hence $\eta = \frac{B}{1+B}$ is the optimal solution. In the optimal π,

$$\bar{R}_1 = (\eta_H + A)\bar{F}(\eta_H) + \frac{AB}{B+1}F(\eta_H) \leq \left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H) + \frac{AB}{B+1}.$$

Hence, the solution is sub-optimal.

- *Case 2: $\eta_L \geq \eta_H$ and $\eta_H < p_G - A$:* If $\eta_L \geq \eta_H$, then $U_1(\eta_H) = U_2(\eta_H) \geq U_2(\eta_L) = 0$ since $U_2(\eta)$ is decreasing in η. Then, $U_1(\eta_H) \geq U_1(p_G - A) = 0$. Since $U_1(\eta)$ is increasing in η, $p_G - A \leq \eta_H$. Therefore, this case is not possible.

- *Case 3: $\eta_L \leq \eta_H$ and $\eta_H > p_G - A$:* If $\eta_L \leq \eta_H$, then $U_1(\eta_H) = U_2(\eta_H) \leq U_2(\eta_L) = 0$ since $U_2(\eta)$ is decreasing in η. Then, $U_1(p_G - A) = 0 \geq U_1(\eta_H)$. Since $U_1(\eta)$ is increasing in η, $p_G - A \geq \eta_H$. Therefore, this case is not possible.

- *Case 4: $\eta_L \le \eta_H$ and $\eta_H \le p_G - A$:* The revenue function for this case is

$$\begin{aligned}\bar{R}_4 = &[\eta_H + (\eta_H - \eta_L)(-\pi + B - B\pi) + A]\,\bar{F}(\eta_H + (\eta_H - \eta_L)(-\pi + B - B\pi)) \\ &+ [\pi\eta_L - B(1-\pi)\eta_L + A\pi]\,F(\eta_L).\end{aligned}$$

Let $\bar{R}_4^{\eta_L}$ be the revenue function for a fixed η_L value and $(\eta_H^{\eta_L}, \pi^{\eta_L})$ be the optimal solution to the problem

$$\underset{\eta_H,\pi}{\text{maximize}} \quad R_4^{\eta_L}(\eta_H, \pi) \tag{14}$$

$$\text{subject to} \quad \eta_H \ge \eta_L,\ \pi \le \frac{B}{1+B},\ \pi \ge 0. \tag{15}$$

We can easily show that $\pi^{\eta_L} = \frac{B}{1+B}$. Suppose $\pi^{\eta_L} < \frac{B}{1+B}$ with $\eta_H^{\eta_L} = \eta_L$. Then, $\bar{R}_4^{\eta_L}(\eta_L, \pi^{\eta_L}) = (\eta_L + A)\,\bar{F}(\eta_L) + [(\pi^{\eta_L} - B + B\pi^{\eta_L})\eta_L + A\pi^{\eta_L}]\,F(\eta_L)$. This function is increasing in π. Hence η_L is not the optimal solution, contradiction. Similarly, suppose $\pi^{\eta_L} < \frac{B}{1+B}$ with $\eta_H^{\eta_L} > \eta_L$. However, the solution can be improved by decreasing η_H and increasing π simultaneously, contradiction. Hence, $\pi^{\eta_L} = \frac{B}{1+B}$. Then

$$\begin{aligned}R_4^{\eta_L} &= max\left\{\max_{\eta_H}(\eta_H + A)\,\bar{F}(\eta_H) + \frac{AB}{1+B}F(\eta_L) \text{ s.t. } \eta_H \ge \eta_L, \left(\eta_L + \frac{A}{1+B}\right)\bar{F}(\eta_L) + \frac{AB}{1+B}\right\} \\ &\le \max_{\eta_H}\left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H) + \frac{AB}{1+B} \text{ s.t. } \eta_H \ge \eta_L.\end{aligned}$$

The solution of (14)–(15) is bounded above by the problem

$$\underset{\eta_H}{\text{maximize}}\left(\eta_H + \frac{A}{1+B}\right)\bar{F}(\eta_H) + \frac{AB}{1+B} \text{ s.t. } \eta_H \ge \eta_L.$$

i.e., $\bar{R}_4 = \underset{\eta_L}{max}\bar{R}_4^{\eta_L}$ is bounded above by R_1. Hence the solution under $\eta_L \le \eta_H$ and $\eta_H \le p_G - A$ is sub-optimal.

Therefore, $\pi \le \dfrac{B}{1+B}$ is sub-optimal.

Proof of Proposition 4 We start with the following lemma.

LEMMA 2. *For any given $(p_G, p_2, \ldots, p_N, \boldsymbol{\pi})$ that satisfies $p_2 \ge p_3 \ge \ldots \ge p_N \ge 0$, $1^T\boldsymbol{\pi} = 1$, $\pi_N \ge \dfrac{B}{1+B}$, and $\boldsymbol{\pi} \ge 0$, there exist $\eta_1 \ge \eta_2 \ge \ldots \ge \eta_N$ such that*

$$\mathcal{S}_G = \{\eta | \eta \ge \eta_1\} \text{ and } \mathcal{S}_{BE}^i = \{\eta | \eta_i \le \eta \le \eta_{i-1}\},\ i = 2, 3, \ldots, N,$$

and similarly, for any given $(\eta_1, \eta_2, \ldots, \eta_N \ge 0, \boldsymbol{\pi})$ that satisfies $\eta_1 \ge \eta_2 \ge \ldots \ge \eta_N$, $1^T\boldsymbol{\pi} = 1$, $\pi_N \ge \dfrac{B}{1+B}$, and $\boldsymbol{\pi} \ge 0$, there exists a set of price levels p_G and $p_2 \ge p_3 \ge \ldots \ge p_N \ge 0$.

Assuming $\eta_1 \ge \ldots \ge \eta_N$, the revenue function of the SP is

$$\begin{aligned}R &= p_G\bar{F}(\eta_1) + \bar{p}_2(F(\eta_1) - F(\eta_2)) + \ldots + \bar{p}_N(F(\eta_{N-1}) - F(\eta_N)) \\ &= (p_G - \bar{p}_2)\bar{F}(\eta_1) + (\bar{p}_2 - \bar{p}_3)\bar{F}(\eta_2) \ldots + (\bar{p}_{N-1} - \bar{p}_N)\bar{F}(\eta_{N-1}) + \bar{p}_N\bar{F}(\eta_N) \\ &= \sum_{i=1}^{N-1}[\pi_i(A + \eta_i) + B\eta_i\pi_i]\,\bar{F}(\eta_i) + [\pi_N(A + \eta_N) - B\eta_N(1 - \pi_N)]\,\bar{F}(\eta_N).\end{aligned}$$

Therefore, the revenue maximization problem becomes

$$\underset{\boldsymbol{\eta},\boldsymbol{\pi}}{\text{maximize}} \quad \sum_{i=1}^{N-1}\pi_i\,[(A + \eta_i) + B\eta_i]\,\bar{F}(\eta_i) + [\pi_N(A + \eta_N) - B\eta_N(1 - \pi_N)]\,\bar{F}(\eta_N) \tag{16}$$

$$\text{subject to} \quad \eta_1 \ge \eta_2 \ge \ldots \ge \eta_N,\ \pi_N \ge \frac{B}{1+B},\ 1^T\boldsymbol{\pi} = 1,\ \boldsymbol{\pi} \ge 0. \tag{17}$$

LEMMA 3. *There exists an optimal solution to the problem above such that at most one of the optimal* $(\pi_1, \pi_2, \ldots, \pi_{N-1})$ *values is nonnegative.*

From Lemma 3, the problem can be simplified to

$$\begin{aligned} \underset{\eta_H, \eta_L, \pi}{\text{maximize}} \quad & (1-\pi)\left[(A+\eta_H)+B\eta_H\right]\bar{F}(\eta_H)+\left[\pi(A+\eta_L)-B\eta_L(1-\pi)\right]\bar{F}(\eta_L) \\ \text{subject to} \quad & \eta_H \geq \eta_L,\ \pi \geq \frac{B}{1+B},\ \pi \leq 1. \end{aligned}$$

which is equivalent to the two-price-level problem.

Proof of Proposition 5 $\Rightarrow$: Assume there are G and BE services in the optimal solution and $p^* \leq v_n$. If $p^* \leq v_n$, then $p^* = v_n$ by the optimality of p^*, which implies there is no customer type that chooses no-buy option. If offering G and BE services together generates more revenue, then there is at least one customer type that chooses BE over G.

When there is only G service, the optimal revenue is $R_1 = \sum_{i=1}^{n} \lambda_i v_n$. When there are two services, the optimal revenue is $R_2 = \sum_{i \in S_1} \lambda_i p_G + \sum_{i \in S_2} \lambda_i \bar{p}_{[i]}$, where S_1 is the set of customer types that chooses G and S_2 is the set of customer types that chooses BE ($S_1 \cap S_2 = \emptyset$ and $S_1 \cup S_2 = \{1, 2, \ldots, n\}$). p_G is the optimal price for G service, which implies $p_G = v_M$ where $M = \max\{i | i \in S_1\}$, and $\bar{p}_{[i]} = \sum_{j=[i]}^{N} \pi_j p_j$ which is the payment of customer type i with her optimal bid value $p_{[i]}$.

Now we will show that $\bar{p}_{[i]} < p_G = v_M\ \forall i \in S_2$. Suppose $\bar{p}_{[k]} \geq p_G = v_M$, $k \in S_2$. Since $k \in S_2$, $\bar{\pi}_{[k]} v_k - (1 - \bar{\pi}_{[k]})\kappa_k - \bar{p}_{[k]} \geq v_k - v_M$, where $\bar{\pi}_{[k]} = \sum_{j=[k]}^{N} \pi_j$. If $\bar{p}_{[k]} \geq v_M$, then $\bar{\pi}_{[k]} v_k - (1 - \bar{\pi}_{[k]})\kappa_k \geq v_k$, which is not possible since $\bar{\pi}_{[k]} < 1$. Therefore, $\bar{p}_{[i]} < p_G = v_M\ \forall i \in S_2$.

$$R_2 = \sum_{i \in S_1} \lambda_i p_G + \sum_{i \in S_2} \lambda_i \bar{p}_{[i]} < \sum_{i=1}^{n} \lambda_i p_G \leq \sum_{i=1}^{n} \lambda_i p^*,$$

where the second inequality comes from the optimality of p^* in one product case. Therefore, offering G and BE services together does not generate more revenue than offering only G service. Contradiction.

$\Leftarrow$: Let $H = \underset{1 \leq i \leq n}{\arg\min}\{p^* \geq v_i\}$. From the optimality of p^*, $v_H = p^*$, and the set of customer types $\{1, 2, \ldots, H\}$ choose G. If we offer a BE service such that no customer types from the set $\{1, 2, \ldots, H\}$ prefer BE and at least one customer type from the set $\{H+1, H+2, \ldots, n\}$ chooses BE, then the revenue generated by G and BE services together becomes higher than that of G service only.

Set $\pi = \dfrac{\kappa_H - \kappa_{H+1}}{v_H - v_{H+1} + \kappa_H - \kappa_{H+1}}$, $p_2 = \dfrac{\kappa_H v_{H+1} - \kappa_{H+1} v_H}{\kappa_H - \kappa_{H+1}}$ and $p_1 = \infty$. Then,

$$v_i - p^* \geq \pi v_i - (1-\pi)\kappa_i - \pi p_2 \ \text{ for } i \leq H \text{ and } \pi v_{H+1} - (1-\pi)\kappa_{H+1} - \pi p_2 \geq 0,$$

which implies all customer types $i \leq H$ choose G and customer type $H+1$ chooses BE service.

Proof of Proposition 6 Let $p_{[i]}$ be the optimal bid for customer type i. Therefore, customer type i either makes a bid of $p_{[i]}$ or leaves the system with no purchase. Clearly, customers with high valuations prefer bidding higher, that is, $p_{[i]}$ is non-increasing in i. Let $s \in \{1, 2, \ldots, n\}$ be the highest customer index that makes a bid, which is determined by $\boldsymbol{p}$ and $\boldsymbol{\pi}$. Therefore, s is not a decision variable. However, an alternative way to solve the problem is to find the optimal $\boldsymbol{p}$ and $\boldsymbol{\pi}$ for any possible s value, and then choose the s that

generates the maximum revenue. Now we characterize and solve the revenue maximization problem for a given s value.

For any $k \in \{1,2,\ldots,n-[i]\}$, type i customer prefers bidding $p_{[i]}$ over $p_{[i]+k}$ if

$$\begin{aligned} & U_2(v_i,\kappa_i,p_{[i]}) \geq U_2(v_i,\kappa_i,p_{[i]+k}) \\ \bar{\pi}_{[i]}v_i - \kappa_i(1-\bar{\pi}_{[i]}) - \bar{p}_{[i]} &\geq \bar{\pi}_{[i]+k}v_i - \kappa_i(1-\bar{\pi}_{[i]+k}) - \bar{p}_{[i]+k} \\ (\bar{\pi}_{[i]} - \bar{\pi}_{[i]+k})(v_i+\kappa_i) &\geq \bar{p}_{[i]} - \bar{p}_{[i]+k}, \qquad i=1,\ldots,s \end{aligned} \tag{18}$$

and prefers bidding $p_{[i]}$ over $p_{[i]-k}$ if

$$\begin{aligned} & U_2(v_i,\kappa_i,p_{[i]-k}) \leq U_2(v_i,\kappa_i,p_{[i]}) \\ \Leftrightarrow\ \bar{\pi}_{[i]-k}v_i - \kappa_i(1-\bar{\pi}_{[i]-k}) - \bar{p}_{[i]-k} &\leq \bar{\pi}_{[i]}v_i - \kappa_i(1-\bar{\pi}_{[i]}) - \bar{p}_{[i]} \\ \Leftrightarrow\ (\bar{\pi}_{[i]-k} - \bar{\pi}_{[i]})(v_i+\kappa_i) &\leq \bar{p}_{[i]-k} - \bar{p}_{[i]}. \qquad i=1,\ldots,s \end{aligned} \tag{19}$$

Lastly, type i customer prefers bidding $p_{[i]}$ over no bidding (i.e., leaving the system with no purchase) if

$$\bar{\pi}_{[i]}v_i - \kappa_i(1-\bar{\pi}_{[i]}) - \bar{p}_{[i]} \geq 0. \qquad i=1,\ldots,s \tag{20}$$

LEMMA 4. *$\bar{\pi}_{[i]}v_i - \kappa_i(1-\bar{\pi}_{[i]})$ is decreasing in i ($i \leq s$).*

LEMMA 5. *(20) can be simplified to*

$$\bar{\pi}_{[s]}v_s - \kappa_s(1-\bar{\pi}_{[s]}) - \bar{p}_{[s]} \geq 0. \tag{21}$$

For a given s, the objective function for the SP is

$$R_s = \sum_{i=1}^{s} \lambda_i \bar{p}_{[i]} = \sum_{j=1}^{s-1}\sum_{i=1}^{j} \lambda_i(\bar{p}_{[j]} - \bar{p}_{[j+1]}) + \sum_{i=1}^{s} \lambda_i \bar{p}_{[s]}.$$

Therefore, the revenue maximization problem can be written as

$$\underset{\boldsymbol{\pi},\boldsymbol{p}}{\text{maximize}} \qquad R_s = \sum_{j=1}^{s-1}\sum_{i=1}^{j} \lambda_i(\bar{p}_{[j]} - \bar{p}_{[j+1]}) + \sum_{i=1}^{s} \lambda_i \bar{p}_{[s]} \tag{22}$$

$$\text{subject to} \qquad (\bar{\pi}_{[i]} - \bar{\pi}_{[i]+k})(v_i+\kappa_i) \geq \bar{p}_{[i]} - \bar{p}_{[i]+k} \qquad i=1,\ldots,s;\ k=1,2,\ldots,N-[i] \tag{23}$$

$$(\bar{\pi}_{[i]-k} - \bar{\pi}_{[i]})(v_i+\kappa_i) \leq \bar{p}_{[i]-k} - \bar{p}_{[i]} \qquad i=1,\ldots,s;\ k=1,2,\ldots,[i]-1 \tag{24}$$

$$\bar{\pi}_{[s]}v_s - \kappa_s(1-\bar{\pi}_{[s]}) - \bar{p}_{[s]} \geq 0 \tag{25}$$

$$p_1 \geq p_2 \geq \ldots \geq p_N \geq 0 \tag{26}$$

$$\mathbf{1}^T\boldsymbol{\pi} = 1,\ \boldsymbol{\pi} \geq 0. \tag{27}$$

LEMMA 6. *Let $(\boldsymbol{\pi}^*,\boldsymbol{p}^*)$ denote an optimal solution to the problem above. Then,*

$$\begin{aligned} p^*_{[1]} = p^*_{[1]+1} = \ldots = p^*_{[2]-1} &= v_1+\kappa_1, \\ p^*_{[2]} = p^*_{[2]+1} = \ldots = p^*_{[3]-1} &= v_2+\kappa_2, \\ &\vdots \\ p^*_{[s-1]} = p^*_{[s-1]+1} = \ldots = p^*_{[s]-1} &= v_{s-1}+\kappa_{s-1}, \\ p^*_{[s]} = p^*_{[s]+1} = \ldots = p^*_N &= v_s + \kappa_s - \frac{\kappa_s}{\bar{\pi}^*_{[s]}}. \end{aligned} \tag{28}$$

Lemma 6 simplifies the revenue maximization problem to

$$\begin{aligned} \underset{\boldsymbol{\pi}}{\text{maximize}} \quad & R_s = \sum_{j=1}^{s-1}\sum_{i=1}^{j}\lambda_i(\bar{\pi}_{[j]}-\bar{\pi}_{[j+1]})(v_j+\kappa_j)+\sum_{i=1}^{s}\lambda_i\left[\bar{\pi}_{[s]}(v_s+\kappa_s)-\kappa_s\right] \\ \text{subject to} \quad & \bar{\pi}_{[s]} \geq \frac{\kappa_s}{v_s+\kappa_s},\ \mathbf{1}^T\boldsymbol{\pi}=1,\ \boldsymbol{\pi}\geq 0. \end{aligned}$$

This problem has N decision variables. It can be simplified further using the following change of variables

$$\begin{aligned} &\alpha_0 = 1-\bar{\pi}_{[1]} \\ &\alpha_i = \bar{\pi}_{[i]}-\bar{\pi}_{[i+1]} && i=1,2,\ldots,s-1 \\ &\alpha_s = \bar{\pi}_{[s]} \end{aligned}$$

where α_i can be interpreted as the time that the price level is equal to $v_i+\kappa_i$ for $i=1,2,\ldots,s-1$, α_0 as the time that the price level is above $v_1+\kappa_1$ and α_s as the time that the price level is at its minimum.

Therefore, the problem becomes

$$\underset{\alpha_0,\alpha_1,\ldots,\alpha_s}{\text{maximize}} \quad R_s = \sum_{j=1}^{s}\sum_{i=1}^{j}\lambda_i\alpha_j(v_j+\kappa_j)-\sum_{i=1}^{s}\lambda_i\kappa_s \tag{29}$$

$$\text{subject to} \quad \alpha_s \geq \frac{\kappa_s}{v_s+\kappa_s},\ \alpha_0+\alpha_1+\ldots+\alpha_s=1,\ \alpha_0,\alpha_1,\ldots,\alpha_s\geq 0. \tag{30}$$

LEMMA 7. *Let* $(\alpha_0^*,\alpha_1^*,\ldots,\alpha_s^*)$ *denote the optimal solution to (29)–(30) and* $k=\underset{j\in\{1,\ldots,s\}}{argmax}\{(v_j+\kappa_j)\sum_{i=1}^{j}\lambda_i\}$. *If* $k=s$, $\alpha_0^*=\alpha_1^*=\ldots=\alpha_{s-1}^*=0,\alpha_s^*=1$, *else* $\alpha_0^*=\ldots=\alpha_{k-1}^*=\alpha_{k+1}^*=\ldots=\alpha_{s-1}^*=0,\alpha_k^*=\frac{v_s}{v_s+\kappa_s},\alpha_s^*=\frac{\kappa_s}{v_s+\kappa_s}$.

Lemma 7 shows that the SP offers at most two price levels, and there is no price level higher than the bid of the highest value customer. Since this result holds for any s, it also holds for the optimal s. Therefore, the optimal solution has at most two price levels.

Proof of Proposition 7 The proposition is equivalent to the following statement.

Let Π_2 be optimal revenue that the SP achieves by offering at most two price levels:

$$\Pi_2 = \max_{1\leq s\leq n} R_s,$$

and Π_1 be the optimal revenue by offering only one price level:

$$\Pi_1 = v_{k^*}\sum_{i=1}^{k^*}\lambda_i$$

where $k^* = \underset{j\in\{1,2,\ldots,n\}}{\arg\max}\left\{v_i\sum_{i=1}^{j}\lambda_i\right\}$. Then, $\Pi_2=\Pi_1$.

If $k^*=n$, i.e., all customers are served with the price v_n, degrading the service for some customers would not increase the revenue, therefore, $k^*<n$ is the first condition to offer two price levels. We need to find two indexes, $\overline{k}$ and $\underline{k}$, the high-level threshold and low-level threshold, respectively, such that customer types $\{1,2,\ldots,\overline{k}\}$ bid high price level, $\{\overline{k}+1,\overline{k}+2,\ldots,\underline{k}\}$ bid low price level, and $\{\underline{k}+1,\underline{k}+2,\ldots,n\}$ leave the system ($\overline{k}\leq\underline{k}\leq n$). Thus, the optimal solution for the two-price case can be written as

$$\Pi_2 = (1-\alpha_{\underline{k}})(v_{\overline{k}}+\kappa_{\overline{k}})\sum_{i=1}^{\overline{k}}\lambda_i+\alpha_{\underline{k}}(v_{\underline{k}}+\kappa_{\underline{k}})\sum_{i=1}^{\underline{k}}\lambda_i-\kappa_{\underline{k}}\sum_{i=1}^{\underline{k}}\lambda_i,$$

Since $\alpha_{\underline{k}} = \frac{\kappa_{\underline{k}}}{v_{\underline{k}}+\kappa_{\underline{k}}}$ from Lemma 7, the optimal revenue for two-price level case becomes

$$\Pi_2 = \frac{v_{\underline{k}}}{v_{\underline{k}} + \kappa_{\underline{k}}} (v_{\overline{k}} + \kappa_{\overline{k}}) \sum_{i=1}^{\overline{k}} \lambda_i.$$

Next we need to find $\overline{k}$ and $\underline{k}$ such that $\Pi_2 > \Pi_1$. Note that Π_2 is decreasing in $\underline{k}$. Therefore, $\underline{k} = \overline{k}$. However, this means that no customer type bids low price level, which is equivalent to one price level solution. Therefore, $\Pi_2 = \Pi_1$.

References

Abhishek, V., I. A. Kash, P. Key. 2012. Fixed and market pricing for cloud services. *arXiv preprint arXiv:1201.5621* .

Afèche, P. 2013. Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management* **15**(3) 423–443.

Afèche, P., H. Mendelson. 2004. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science* **50**(7) 869–882.

Afèche, P., M. Pavlin. 2015. Optimal price/lead-time menus for queues with customer choice: segmentation, pooling, strategic delay. *Management Science (forthcoming)* .

Agmon Ben-Yehuda, O., M. Ben-Yehuda, A. Schuster, D. Tsafrir. 2011. Deconstructing amazon ec2 spot instance pricing. *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*. IEEE, 304–311.

Anderson, E. T., J. D. Dana, Jr. 2009. When is price discrimination profitable? *Management Science* **55**(6) 980–989.

Armbrust, M., A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, et al. 2010. A view of cloud computing. *Communications of the ACM* **53**(4) 50–58.

Baron, O. 2003. Pricing and admission control for shared computer services using the token bucket mechanism. Ph.D. thesis, Massachusetts Institute of Technology.

Borgs, C., O. Candogan, J. Chayes, I. Lobel, H. Nazerzadeh. 2014. Optimal multiperiod pricing with service guarantees. *Management Science* **60**(7) 1792–1811.

Deneckere, R. J., P. R. McAfee. 1996. Damaged goods. *Journal of Economics & Management Strategy* **5**(2) 149–174.

Katta, A., J. Sethuraman. 2005. Pricing strategies and service differentiation in queues – a profit maximization perspective. Tech. rep., Computational Optimization Research Center, Columbia University. TR-2005-04.

Maglaras, C., J. Yao, A. Zeevi. 2015. Optimal price and delay differentiation in queueing systems. *Management Science (forthcoming)* .

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* **49**(8) 1018–1038.

Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research* **53**(2) 242–262.

McAfee, R. P. 2007. Pricing damaged goods. Economics Discussion Paper 2007-2, Kiel Institute for the World Economy. URL `http://www.economics-ejournal.org/economics/discussionpapers/2007-2`.

Mendelson, H. 1985. Pricing computer services: queueing effects. *Communications of the ACM* **28**(3) 312–321.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations Research* **38**(5) 870–883.

Mitra, D., Q. Wang. 2015. Preservation of best-effort service on the internet in the presence of managed services and usage-generated applications. *Available at SSRN 2587828* .

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.

Nazerzadeh, H., R. S. Randhawa. 2015. Near-optimality of coarse service grades for customer differentiation in queueing systems. *Available at SSRN 2438300* .

Savin, S. V., M. A. Cohen, N. Gans, Z. Katalan. 2005. Capacity management in rental businesses with two customer bases. *Operations Research* **53**(4) 617–631.

Xu, H., B. Li. 2013. A study of pricing for cloud resources. *ACM SIGMETRICS Performance Evaluation Review* **40**(4) 3–12.

Appendix. Online Supplement. Additional Proofs

Proof of Lemma 1 Bidding a value between two price levels is the same as bidding the lower price level amount in terms of availability of the product and payment amount, therefore sub-optimal.

Proof of Lemma 2 $\mathcal{S}^i_{BE}$ contains all η values that satisfy the following constraints

$$U_2(\eta, p_i) \geq U_2(\eta, p_{i+j}), j = 1, \ldots, N-i \tag{31}$$

$$U_2(\eta, p_i) \geq 0, \tag{32}$$

$$U_2(\eta, p_i) \geq U_2(\eta, p_{i-j}), j = 1, \ldots, i-2 \tag{33}$$

$$U_2(\eta, p_i) \geq U_1(\eta). \tag{34}$$

There exists an $\underline{\eta}_i$ such that any $\eta \geq \underline{\eta}_i$ satisfies (31) and (32). Similarly, there exists an $\overline{\eta}_i$ such that any $\eta \leq \overline{\eta}_i$ satisfies (33) and (34). Hence, $\mathcal{S}^i_{BE} = \{\eta | \underline{\eta}_i \leq \eta \leq \overline{\eta}_i\}$. For the rest of the analysis, we assume $\underline{\eta}_i \leq \overline{\eta}_i$ for $i = 2, 3, \ldots, N$.

Next, we will show that $\overline{\eta}_{i+1} \leq \underline{\eta}_i$ for $i = 2, 3, \ldots, N-1$. Suppose there exists an i such that $\overline{\eta}_{i+1} > \underline{\eta}_i$. Then, since $\bar{\pi}_i \geq \bar{\pi}_{i+1} \geq 0$, $U_2(\overline{\eta}_{i+1}, p_i) > U_2(\overline{\eta}_{i+1}, p_{i+1})$. However, from 33, $U_2(\overline{\eta}_{i+1}, p_{i+1}) \geq U_2(\overline{\eta}_{i+1}, p_i)$. Contradiction. Therefore, $\overline{\eta}_{i+1} \leq \underline{\eta}_i$ for $i = 2, 3, \ldots, N-1$.

Now, we will find conditions for $\underline{\eta}_i, \overline{\eta}_i$ for $i = 2, 3, \ldots, N$. First, we show that $U_2(\underline{\eta}_i, p_i) = U_2(\underline{\eta}_i, p_{i+1})$ for $i = 2, 3, \ldots, N-1$. Suppose it is not true, which implies $U_2(\underline{\eta}_i, p_i) = U_2(\underline{\eta}_i, p_{i+k}) > U_2(\underline{\eta}_i, p_{i+1})$ for some $k > 1$. Moreover, $U_2(\overline{\eta}_{i+1}, p_{i+1}) \geq U_2(\overline{\eta}_{i+1}, p_{i+k})$. Since $\underline{\eta}_i \geq \overline{\eta}_{i+1}$ and $U_2'(\eta, p_{i+1}) \geq U_2'(\eta, p_{i+k})$, $U_2(\underline{\eta}_i, p_{i+1}) \geq U_2(\underline{\eta}_i, p_{i+k})$. Contradiction. Therefore, $U_2(\underline{\eta}_i, p_i) = U_2(\underline{\eta}_i, p_{i+1})$. For $i = N$, $U_2(\underline{\eta}_N, p_N) = 0$. Second we show that $U_2(\overline{\eta}_i, p_i) = U_2(\overline{\eta}_i, p_{i-1})$ for $i = 3, 4, \ldots, N$. Suppose not true, which implies $U_2(\overline{\eta}_i, p_i) = U_2(\overline{\eta}_i, p_{i-k}) > U_2(\overline{\eta}_i, p_{i-1})$ for some $k > 1$. Moreover, $U_2(\underline{\eta}_{i-1}, p_{i-1}) \geq U_2(\underline{\eta}_{i-1}, p_{i-k})$. Since $\underline{\eta}_{i-1} \geq \overline{\eta}_i$ and $U_2'(\eta, p_{i-1}) \geq U_2'(\eta, p_{i-k})$, $U_2(\overline{\eta}_i, p_{i-1}) \geq U_2(\overline{\eta}_i, p_{i-k})$. Contradiction. Therefore, $U_2(\overline{\eta}_i, p_i) = U_2(\overline{\eta}_i, p_{i-1})$. For $i = 2$, $U_2(\overline{\eta}_2, p_2) = U_1(\overline{\eta}_2)$. From the two conditions on $\underline{\eta}_i$ and $\overline{\eta}_i$, we reach $\underline{\eta}_{i-1} = \overline{\eta}_i$ for $i = 3, 4, \ldots, N$.

Next step is to rename the boundaries. Let $\eta_i = \underline{\eta}_i$ for $i = 2, 3, \ldots, n$ and $\eta_1 = \overline{\eta}_2$. This concludes the first part of the proposition.

For any given $(\eta_1, \eta_2, \ldots, \eta_N, \boldsymbol{\pi})$ that satisfies $\eta_1 \geq \eta_2 \geq \ldots \geq \eta_N$, $1^T\boldsymbol{\pi} = 1$, $\pi_N \geq \dfrac{B}{1+B}$, and $\boldsymbol{\pi} \geq 0$, the following prices satisfy $p_2 \geq p_3 \geq \ldots \geq p_n \geq 0$ and they are aligned with $\mathcal{S}_G$ and $\mathcal{S}^i_{BE}$, $i = 2, 3, \ldots, N$:

$$\begin{aligned} p_G &= A + \eta_1, \\ p_i &= A + \eta_i + B\eta_i, && i = 2, 3, \ldots, N-1, \\ p_N &= A + \eta_N + B\eta_N - \frac{B\eta_N}{\pi_N}. \end{aligned}$$

Proof of Lemma 3 Let $(\boldsymbol{\eta}^*, \boldsymbol{\pi}^*)$ be an optimal solution. Assume that there are two π^*_i $(i = 1, 2, \ldots, N-1)$ values such that $\pi^*_j > 0$, $\pi^*_k > 0$, and all others are equal to 0. Without loss of generality, $j < k$ which implies $\eta^*_j \geq \eta^*_k$. If $\eta^*_j = \eta^*_k$, then another optimal solution would be $\pi^*_j := \pi^*_j + \pi^*_k$ and $\pi^*_k := 0$, which has only one nonnegative π_i value for $i = 1, \ldots, N-1$. If $\eta^*_j > \eta^*_k$, then there are three possible cases:

Case 1: $[(A + \eta^*_j) + B\eta^*_j] = [(A + \eta^*_k) + B\eta^*_k]$: $\pi^*_j := \pi^*_j + \pi^*_k$ and $\pi^*_k := 0$ is another optimal solution where at most one π value is nonnegative.

Case 2: $[(A+\eta_j^*)+B\eta_j^*] > [(A+\eta_k^*)+B\eta_k^*]$: $\pi_j^* := \pi_j^* + \pi_k^*$ and $\pi_k^* := 0$ give a better solution, contradiction.

Case 3: $[(A+\eta_j^*)+B\eta_j^*] < [(A+\eta_k^*)+B\eta_k^*]$: $\pi_j^* := 0$ and $\pi_k^* := \pi_j^* + \pi_k^*$ give a better solution, contradiction.

Therefore, there cannot be two nonnegative π_i values ($i = 1, 2, \ldots, N-1$). Using the same idea, we can generalize the result to more than two nonnegative value case.

Proof of Lemma 4 From (20) and prices being nonnegative,

$$\bar{\pi}_{[i]} \geq \frac{\kappa_i}{v_i + \kappa_i}, \qquad i = 1, \ldots, s$$

and since $\frac{v_i}{\kappa_i}$ is decreasing in i,

$$\frac{\kappa_{i+1}}{v_{i+1} + \kappa_{i+1}} \geq \frac{\kappa_i}{v_i + \kappa_i}. \qquad i = 1, 2, \ldots, s-1.$$

Using these two inequalities and $\bar{\pi}_{[i]}$ being decreasing in i, for any $i = 1, 2, \ldots, s-1$,

$$\begin{aligned}
\bar{\pi}_{[i]} v_i - \kappa_i(1 - \bar{\pi}_{[i]}) &= \bar{\pi}_{[i]}(v_i + \kappa_i) - \kappa_i \\
&= \bar{\pi}_{[i+1]}(v_i + \kappa_i) + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i + \kappa_i) - \kappa_i \\
&= \bar{\pi}_{[i+1]}\left[v_i + \kappa_i - (v_{i+1} + \kappa_{i+1})\right] + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i + \kappa_i) - \kappa_i + \bar{\pi}_{[i+1]}(v_{i+1} + \kappa_{i+1}) \\
&\geq \frac{\kappa_{i+1}}{v_{i+1} + \kappa_{i+1}}(v_i + \kappa_i) - \kappa_{i+1} + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i + \kappa_i) - \kappa_i + \bar{\pi}_{[i+1]}(v_{i+1} + \kappa_{i+1}) \\
&\geq \frac{\kappa_i}{v_i + \kappa_i}(v_i + \kappa_i) - \kappa_{i+1} + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i + \kappa_i) - \kappa_i + \bar{\pi}_{[i+1]}(v_{i+1} + \kappa_{i+1}) \\
&= \bar{\pi}_{[i+1]}(v_{i+1} + \kappa_{i+1}) - \kappa_{i+1} + (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i + \kappa_i) \\
&\geq \bar{\pi}_{[i+1]}(v_{i+1} + \kappa_{i+1}) - \kappa_{i+1} \\
&= \bar{\pi}_{[i+1]} v_{i+1} - \kappa_{i+1}(1 - \bar{\pi}_{[i+1]})
\end{aligned}$$

Proof of Lemma 5 Using Lemma 4, it can easily be shown that

$$\bar{\pi}_{[i+1]}(v_i + \kappa_i) - \kappa_i \geq \bar{\pi}_{[i+1]}(v_{i+1} + \kappa_{i+1}) - \kappa_{i+1}. \qquad i = 1, 2, \ldots, s-1$$

Then, using (18), for any $i = 1, 2, \ldots, s-1$,

$$\begin{aligned}
\bar{\pi}_{[i]} v_i - \kappa_i(1 - \bar{\pi}_{[i]}) - \bar{p}_{[i]} &= \bar{\pi}_{[i]}(v_i + \kappa_i) - \kappa_i - \bar{p}_{[i+1]} - (\bar{p}_{[i]} - \bar{p}_{[i+1]}) \\
&\geq \bar{\pi}_{[i]}(v_i + \kappa_i) - \kappa_i - \bar{p}_{[i+1]} - (\bar{\pi}_{[i]} - \bar{\pi}_{[i+1]})(v_i + \kappa_i) \\
&= \bar{\pi}_{[i+1]}(v_i + \kappa_i) - \kappa_i - \bar{p}_{[i+1]} \\
&\geq \bar{\pi}_{[i+1]}(v_{i+1} + \kappa_{i+1}) - \kappa_{i+1} - \bar{p}_{[i+1]}.
\end{aligned}$$

Therefore,

$$\bar{\pi}_{[1]} v_1 - \kappa_1(1 - \bar{\pi}_{[1]}) - \bar{p}_{[1]} \geq \bar{\pi}_{[2]} v_2 - \kappa_2(1 - \bar{\pi}_{[2]}) - \bar{p}_{[2]} \geq \ldots \geq \bar{\pi}_{[s]} v_s - \kappa_s(1 - \bar{\pi}_{[s]}) - \bar{p}_{[s]} \geq 0.$$

Proof of Lemma 6 First, we need to show that (28) is a feasible solution. Since $v_i + \kappa_i$ is decreasing in i, and (25) implies $\bar{\pi}^*_{[s]} \geq \frac{\kappa_s}{v_s + \kappa_s} > 0$, the solution satisfies (26). (27) is trivially satisfied since (28) does not impose anything on $\boldsymbol{\pi}$ and uses the optimal $\boldsymbol{\pi}^*$, which is also feasible. (25) holds with equality. Trivially (23)

and (24) are also satisfied. Therefore, (28) is a feasible solution. Now we need to show that this solution is optimal. The objective function is equivalent to

$$R_s = \sum_{i=1}^{s} \lambda_i \bar{p}_{[i]},$$

where all $\bar{p}_{[i]}$ variables have nonnegative coefficients. Moreover, (28) provides a solution where all p variables are equal to their upper bounds. Therefore, the solution is an optimal solution.

Proof of Lemma 7 Suppose $k = s$ and $\alpha_s^* = 1 - \varepsilon$ with $\alpha_j^* = \varepsilon$ $(0 \leq j < s, \varepsilon > 0)$. Since $(v_s + \kappa_s) \sum_{i=1}^{s} \lambda_i > (v_j + \kappa_j) \sum_{i=1}^{j} \lambda_i$, $\alpha_s = 1, \alpha_j = 0$ gives a higher objective function value. Contradiction. Similarly, if $k < s$, α_s^* has to be equal to its lower bound in the optimal solution. If $\underset{j \in \{1,\ldots,s\}}{argmax}\{(v_j + \kappa_j) \sum_{i=1}^{j} \lambda_i\}$ is not unique, there are alternative optimal solutions which has a solution given above. This proves that the SP offers at most two price levels. Moreover, since $k > 0$, $\alpha_s^* = 0$, which means the fraction of time the price level is above $v_1 + \kappa_1$ is equal to zero. Therefore, the price never goes beyond the bid of the highest value customer, which is equal to $v_1 + \kappa_1$.