Pricing Models for On-Demand Computing

Ke-Wei Huang[1] and Arun Sundararajan[2]

Working Paper CeDER-05-26, Center for Digital Economy Research
Leonard N. Stern School of Business, New York University

November 2005

Abstract: On-demand computing provides a new way for companies to manage and use their IT infrastructure. This model of corporate computing radically changes the way companies pay for their IT infrastructure, basing it on "pay per use" rather than on the fixed infrastructure investments such companies are accustomed to. A clear theoretical understanding of pricing on-demand computing is thus central to the viability and growth of this nascent industry. We contribute towards such an understanding in this paper by modeling the optimal pricing of on-demand computing while taking four critical factors into account: the costs of deploying IT in-house, the business value of this IT, the scale of the provider's on-demand computing infrastructure, and the variable costs of providing on-demand computing. Three distinct pricing models emerge as optimal among all possible pricing functions for on-demand computing. These models describe when volume discounting, free usage and demand caps should be used to manage demand appropriately and profitably. We also outline a likely path that the transformation towards on-demand computing will follow – under which low-usage customers are targeted initially, followed by a broadening of the market, and finally, a focus on profiting from inducing adoption by high-usage customers – and prescribe how the associated pricing models should evolve appropriately[3].

JEL Codes: D42, D82, L12, L86.

[1] 44 West 4th Street, KMC 8-185, New York, NY 10012. *khuang0@stern.nyu.edu*
[2] 44 West 4th Street, KMC 8-93, New York, NY 10012. *asundara@stern.nyu.edu*
[3] We thank seminar participants at New York University for their feedback. The usual disclaimers apply.

1 Introduction

The emergence of on-demand computing promises to transform the way corporations buy and manage their IT infrastructure. This model of computing, also referred to as "utility computing", or in the specific context of software, as "apps on tap", may shift IT infrastructure from being a fragmented capital asset to being a centralized utility service. The viability of on-demand computing has been facilitated by two related technological developments – grid computing and Web services. Together with the widespread availability of Internet bandwidth, they make it technologically feasible for corporate buyers to "rent" key parts of their IT infrastructure – servers, data storage, isolated software applications, and integrated software solutions, for, among other things, salesforce management, CRM and retail fulfilment – from large-scale utility providers, rather than deploying and running these parts of their infrastructure in-house.

Current spending levels for on-demand computing are still a small fraction of IT corporate budgets; however, this fraction is widely projected as growing rapidly over the next decade. During this nascent stage in its evolution, it is essential for providers of on-demand computing develop and implement pricing and migration strategies that make the transition appropriately gradual and reliable. An inappropriate choice of pricing that is based on usage could either lead to excessive inertia in migration, or alternatively, to excess demand that providers cannot fulfil profitably or scale to meet reliably. Either scenario could easily kill early innovators in on-demand computing.

The importance of a careful and judicious choice of pricing models, critical during this transition in corporate computing, motivates our paper's objective: to develop robust prescriptions for pricing on-demand computing. We identify and model four aspects of corporate IT infrastructure that affect it's pricing:

(1) The cost of buying, deploying and maintaining the infrastructure *in-house*: Since this

option is always a viable substitute for on-demand computing, the magnitude of its cost will play an important part in pricing on-demand computing.

(2) The business value of the infrastructure: Often, the business value of different kinds of IT infrastructure varies widely, based on its functionality and the context in which it is used, and this variation is not always directly related to its cost. Providers of on-demand computing are likely to price based on business value or willingness to pay, while their customers may be accustomed to thinking about paying for the infrastructure based on its cost.

(3) The scale of the provider's on-demand computing infrastructure: The seller's primary instrument to control demand is its pricing model, which needs to be designed to ensure that its infrastructure is not under-utilized, while also maintaining an acceptable level of quality-of-service and reliability, both of which can be compromised by excessive demand. Linking infrastructure scale to pricing is therefore important.

(4) The variable costs of on-demand computing: Even after accounting for the costs of infrastructure, on-demand computing services are not "information goods". Rather, their provision involves non-trivial variable costs that relate to customer service, billing and monitoring. Additionally, adopting customers may bear variable costs of transition, and of usage, that are independent of price.

Our analysis of optimal pricing for on-demand computing incorporates each of these four factors, and yields three different optimal pricing models, each of which characterizes a distinct stage of transition in corporate computing. These models may feature volume discounting, free usage for high-demand customers, and demand caps. The emergence of a range of nonlinearity in the optimal pricing policy validates our approach of choosing to place no restrictions on the choice of pricing function, despite the associated analytical complexity of this approach.

We are also able to characterize which model is suitable based on the stage of evolution of an infrastructure industry towards the on-demand model, and suggest a path of transition

pricing for an IT industry making the transition. In the early stages of transition to on-demand computing, our model prescribes an optimal pricing strategy designed to attract lower-usage customers, by offering demand-capped pricing that draws in those companies who lack the scale to implement their own in-house solutions. In an intermediate stage, as providers gain confidence in the viability of the on-demand computing model, a pricing strategy which features limited free-usage to some higher-usage customers is optimal. This strategy will induce a measured expansion into the middle of the market (which, although seemingly related, is quite different from the "move to the middle" hypothesis of Clemons, Reddi and Rowe, 1995), while continuing to profit from the lower-usage customers attracted early in the process. As the on-demand computing model matures, and the fixed costs of in-house deployment rise as a consequence, pricing is likely to become more "normal": aimed at a broad segment of high-usage customers, with volume discounts and no artificial caps on demand. Our model also suggests that the sectors of the IT industry that make the transition earlier are likely to be those whose impact on an organization's core business value is not as substantial.

Our work extends the literature of nonlinear pricing with positive participation constraints (Lewis and Sappington 1989, Maggi and Rodriguez 1995, and Jullien 2000) and the literature about on-demand computing. The former literature is quite vast, a survey is beyond the scope of our paper, and we therefore focus on placing our work in the latter (and related) literature. Gurnani and Karlapalem (2001) develop a monopoly pricing model to examine the optimal pricing strategies for selling and pay per use licensing of packaged software over Internet. Their main contribution is to show that pay per use is indeed a profitable alternative for software vendors. Snir and Hitt (2003) investigate bidding behavior in early-stage markets for IT services, theorizing that such bidding contains important information about the viability and value of these markets early in their evolution, and finding that services with higher value do indeed attract a larger set of bids. Paleologo (2004) presents a methodology for pricing utility computing

that takes risk into account, and reports on how it improves on simple cost-plus pricing models. Chen and Wu (2004) model a seller's choice of linear usage-based pricing for on-demand computing, and Bhargava and Sundaresan (2004) study how to use contingent auctions to price utility computing. Our paper also adds to a growing literature in information systems that study second-degree pricing discrimination for IT-related goods and services. This literature includes Nault (1997), who analyzes pricing and profitability when an interorganizational system can lead to quality differentiation based on whether or not a good is supported by the IOS, and relates these measures to the optimal design of the IOS for both a monopolist and competing duopolists., Bakos and Brynjolfsson (1999) who show that bundling information goods is often profitable, Geng, Stinchcombe and Whinston (2005) who show that pure bundling may not be optimal when the value each customer places on different goods in the bundle varies sufficiently, Bhargava and Choudhary (2001) who establish conditions under which versioning is optimal, Weber (2001) who analyzes when mixed versioning is optimal, Anand and Aron (2003), who derive a monopolist's optimal group-buying schedule and compare its profits with those that obtain under the more conventional posted-price mechanism, and Sundararajan (2004) who establishes the optimality of fixed-fee pricing for information goods.

Our paper differs from this stream of literature in the following key ways. First, we explicitly model the presence of a "build-your-own" option, which while possibly less relevant for information goods, is clearly a critical aspect of a corporation's choice of IT infrastructure, and consequently, of a seller's pricing model for on-demand computing targeted at such corporations. Second, we explicitly model the effects of the seller's choice of infrastructure levels: in other words, we recognize and incorporate the fact that while their costs at the margin may often be zero, on-demand computing services are not really information goods: increasing supply often necessitates discrete and costly increases in infrastructure. Third, we explicitly model the choice of a pricing function, rather than imposing a linear pricing function ex-ante. Our results

show that the structure of this pricing function can vary significantly at different phases in the industry's evolution, and can include volume discounts, free usage and demand caps, none of which would emerge in an analysis of linear pricing. This highlights the importance of our generalization. Our model is similar in some respects to the model in Sundararajan (2004b), which studies optimal nonlinear pricing in the presence of piracy, since both models study price discrimination in the presence of an outside alternative whose value depends on customer type. What distinguishes our model is that the outside alternative for on-demand computing involves payments for fixed infrastructure, thereby significantly changing the way surplus varies across customer type, and leading to a broader family of pricing functions that emerge as optimal.

We do not explicitly model queuing effects or congestion costs. Beginning with the seminal paper by Mendelson (1985), there is a rich literature in IS that studies the effects of such costs on pricing. These include Mendelson and Whang (1990) who demonstrate that priority pricing for a constrained resource can be expressed in a simple form: a base price for the lowest class plus an increasing priority surcharge; Dewan and Mendelson's (1990) analysis of congestion pricing with general delay costs, Dewan's (1996) discussion of declining computing costs on the tradeoff between capacity costs and user time, Konana, Gupta and Whinston's (2000) model of dynamic priority pricing with congestion premiums, and Afeche and Mendelson (2004) who model interdependent delay costs and consumer value. Granted, congestion costs are certainly pertinent to the provision of on-demand computing, and we do not ignore them entirely. Instead, we translate such effects into a shadow cost imposed by the level of infrastructure, based on an implicit assumption of a constant level of quality of service. By abstracting congestion costs in this way, we can focus on modeling of other aspects of pricing on-demand computing that have not received as much attention in the IS literature.

We have organized the rest of this paper as follows. Section 2 presents our model which characterizes the demand for computing, the cost structure of on-demand and in-house de-

ployment, and formulates the pricing problem for a seller of on-demand computing. Section 3 presents the solution to this problem. It describes three distinct pricing models, each of which is optimal across a different range of combinations of average business value from the service, and the relative fixed and variable costs of in-house and on-demand deployment, and discusses the distinct characteristics of each of these models. Section 4 summarizes our results, discusses a possible path of transition towards on-demand computing and the corresponding pricing models that will accompany it, and outlines directions for future research.

2 Model

This section briefly describes our model of the costs and value from on-demand computing, and concludes with a characterization of the pricing problem that is solved in Section 3.

2.1 Overview

We model a computing or IT infrastructure service, henceforth simply referred to as the *service*, that is used by different customers in varying quantities. These customers are heterogeneous in how much they value a specific number of units q of the service (a couple of examples are provided later in this section that illustrate what these units might be for different kinds of on-demand computing). Specifically, each customer is indexed by a parameter θ that determines the value $u(q, \theta)$ that the customer derives from using q units of the service, and which we assume takes the following functional form:

$$u(q,\theta) = \begin{cases} \alpha[\theta q - \frac{1}{2}q^2], & q \leq \theta \\ \alpha[\frac{1}{2}\theta^2], & q > \theta \end{cases}, \tag{1}$$

where α is a service-specific parameter that models how much business value customers derive from the service, on average. This specification of $u(q, \theta)$ implies that the parameter θ represents the maximum demand a customer indexed by θ has for the service, and, for a fixed value of α,

is also indicative of the marginal *business value* the customer derives from an increase in their demand for the service. We assume that θ is known only to the customer (that is, it cannot be observed by a seller), and that it is uniformly distributed[2] on the interval $[0, 1]$.

To fulfil their demand for the service, customers have two options. The first option is to buy, implement and maintain the infrastructure required to fulfil the service *in-house*, from an established market for in-house fulfilment of the service. A customer who chooses this option incurs a fixed infrastructure cost of F, and this option enables the customer to fulfil their entire computing needs[3].

The second option a customer has is to fulfil all or part of their demand from a monopoly seller of an *on-demand* version of the service. This seller chooses a pricing function $P(q)$ that specifies the price that a customer will pay for their usage of q units of the service. Since the seller cannot observe the type θ of any customer, this pricing function is available to all customers. A special (familiar) example of a pricing function might be a linear price $P(q) = pq$, and a natural question at this stage is whether additional insight can be gained by allowing the seller to choose a more general pricing function. Our results in Section 3 will show that linear pricing is never optimal for the seller: the pricing model chosen will feature increasing volume discounts, demand caps, free usage for high-usage customers, or some combination of these factors.

Our model easily generalizes to a scenario in which there are a few distinct segments in which customers across segments differ in terms of their scale (thus incurring a segment-specific cost of in-house computing F), so long as the seller can identify what segment a customer belongs to,

[2]Our choice of utility function and distribution are for analytical simplicity. The pricing structures we derive are likely to generalize in their structure to a more abstract model, although establishing this will involve substantial additional analytical complexity.

[3]In a different context, this may be interpreted as an "all you can use" price for a tangible version of a digital good now being priced as a service.

but faces some residual incomplete information about specific customers within each segment. For example, a seller may be able to identify whether a potential corporate customer is a small business or a Fortune 500 company, or may have a rough estimate of how many employees or customers a company has; this information is indicative of both the customer's willingness to pay for the service, as well as the scale of costs they incur to deploy in-house computing. So long as a customer's segment can be identified, the seller will simply *independently* design a different pricing schedule for each segment according to our results in Section 3, and offer only that schedule to customers in the segment. To that extent, our assumption of one segment with a common cost of in-house deployment F is without much loss in generality.

The seller deploys a fixed level of infrastructure, represented by the vector $K = (k_1, k_2, ..., k_n)$. The components of infrastructure could include hardware, software licenses, disk storage, customer support infrastructure, administration and maintenance staff, and so on. The fixed cost of this infrastructure is sunk and therefore does not affect the seller's choice of pricing, although the level of K does affect pricing through the constraint it places on a seller's ability to fulfil demand. The seller also incurs a linear variable cost $c \geq 0$ per unit of demand it fulfils from this shared infrastructure. This cost may be incurred due to transaction costs associated with billing, usage monitoring and customer support, or may include costs associated with integrating the service with other parts of the customer's existing in-house IT infrastructure. The seller also guarantees a fixed level of quality-of-service to each of its customers, which restricts the aggregate level of demand that it can fulfil at a choice of infrastructure K to a maximum of $Q(K)$. The function $Q(K)$ is (weakly) increasing in each component of K (a higher level of infrastructure leads to a higher maximum demand). Our specification of infrastructure is analogous to the idea of IOS design used in Nault (1997), where the vector K of infrastructure is similar to his vector $\overrightarrow{x}$ of IT inputs, which are mapped to IOS quality using a production function $\xi(\overrightarrow{x})$ Beyond this, it is not necessary to make any assumptions about its functional

form, and the connection between K and the seller's choice of pricing will be made clear at the beginning of Section 3.

To illustrate the variables of the model further, consider the example of on-demand access to a grid of high-performance servers. The units of q in this example could be the number of processor-hours that the customer uses (this is the basis for Sun Microsystems' pricing of access to their grid, for example). F would represent the cost of acquiring and administering a high-performance server in-house, and $Q(k_1)$ would represent the maximum number of processor hours that, say, a grid with k_1 nodes can support. A second example might be on-demand access to Amazon's retailing platform. The units of q in this second example could be the number of (some standardized) transactions the customer fulfils using Amazon's platform. F would be the cost to the customer of buying, setting up and running the hardware, application and database server software, and other infrastructure required for the in-house fulfilment of online retailing, and $Q(K)$ would represent the maximum number of non-Amazon (partner) retail transactions Amazon.com can support under its current infrastructure level K.

2.2 The trade-off between in-house and on-demand corporate computing

Once the pricing schedule $P(q)$ is set, for a customer of type θ, the surplus from a choice of on-demand deployment is:

$$\max_q [u(q, \theta) - P(q)], \tag{2}$$

and the surplus from a choice of in-house deployment is

$$\left(\max_q [u(q, \theta)] \right) - F. \tag{3}$$

Therefore, customers segment into a maximum of three groups, based on the absolute value of their surplus from each option, as well as the relative values of surplus between options. The first group chooses on-demand computing, and will consist of those customers whose type θ is

such that their net value from on-demand computing is positive, and is higher than their net value from in-house deployment:

$$\max_q[u(q,\theta) - P(q)] \geq 0. \tag{4}$$

$$\max_q[u(q,\theta) - P(q)] \geq \left(\max_q[u(q,\theta)]\right) - F. \tag{5}$$

For simplicity, we assume that indifferent customers choose on-demand computing. The second group chooses in-house deployment, and will consist of those customers whose type θ is such that their net value from in-house deployment is positive, and is higher than their net value from in-house deployment:

$$\left(\max_q[u(q,\theta)]\right) - F \geq 0. \tag{6}$$

$$\left(\max_q[u(q,\theta)]\right) - F > \max_q[u(q,\theta) - P(q)]. \tag{7}$$

The third group of customers will be those whose net value from either option is negative, and who therefore do not deploy the service at all.

The seller's pricing problem is to choose the pricing function $P(q)$ that maximizes its profits, given that its demand will be from the set of customers whose type θ satisfies (4) and (5), that each customer in this set will generate demand according to (2), and that total demand fulfilled cannot exceed $Q(K)$. We transform this formulation into a simpler one to facilitate solving it, and the mathematical details of this transformation are presented in Appendix A.

3 Pricing models for on-demand computing

In this section, we present the seller's optimal pricing schedule. It takes the form of being one of three different pricing models. We describe each model and relate its choice to the seller's cost structure, infrastructure choices, and the business value customers generate from their usage of the service. We present our results as a sequence of three pricing models for

clarity, and to be able to highlight the features of each that are relevance to pricing in practice. However, rather than being specific structures (or functional forms) we have exogenously chosen to study, these are the pricing functions that our analysis establishes as optimal among *all possible pricing functions*, and across all feasible combinations of the model's parameters. That is, we analyze all possible combinations of parameters, and derive the optimal pricing function for *each* combination of parameters. To make our exposition clearer and easier to understand, and towards trying to make sure that the implications of our results do not get buried in the complexity of the math that our analysis entails, we have then partitioned the parameter space (as summarized in Figure 1) based on structural similarities we identify within different optimal pricing functions. Appendix B presents the mathematical details of our analysis and contains all of our proofs.

3.1 A preliminary result: relating infrastructure levels to pricing

Our first lemma is an intermediate result that relates the seller's choice of infrastructure K to their pricing problem.

Lemma 1 *The seller's pricing problem with the constraint $Q(K)$ on demand is equivalent to the same pricing problem with no constraint on demand, but with linear variable costs $c+\lambda(K)$, where $\lambda(K)$ measures the "shadow" cost of using one unit of demand from $Q(K)$, and is (weakly) decreasing in each component of K.*

This result allows us to view the seller's infrastructure constraint as being equivalent to bearing an additional (variable) shadow cost $\lambda(K)$ that is linear in total demand. There is always an infrastructure level for which this cost is zero. This is at a point where the seller can fulfil its unconstrained profit-maximizing level of the demand. None of the results that follow (for instance, the presence of demand caps) rely critically on the presence of this "capacity"

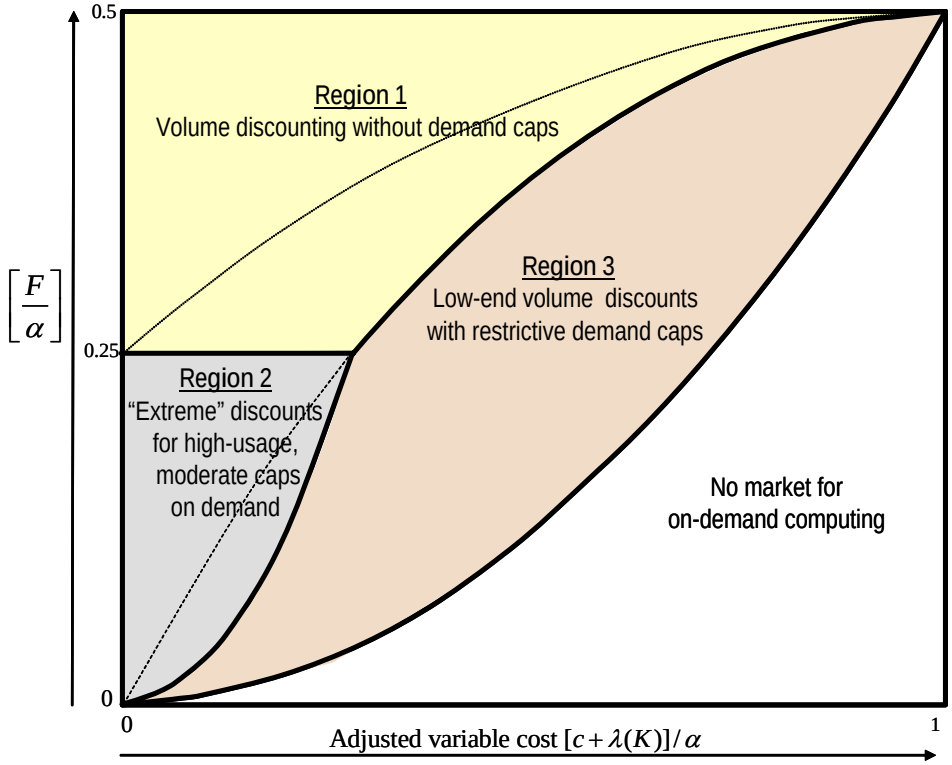

Figure 1: Summarizes the three pricing models for different average business value, cost and infrastructure levels. Each of these models is discussed in more detail later in this section. The analytical details of each of the curves that define this partition are summarized in Table 1.

constraint – rather, it simply allows infrastructure K to affect pricing, if in fact the level of K constrains demand.

Our next three results describe the seller's optimal pricing schedule under different combinations of $c, \lambda(K)$, F, and α. The regions of the parameter space under which each is optimal are summarized in Figure 1, and the exact expressions defining these regions are summarized in Table 1. Notice that these are exhaustive, accommodating all feasible positive combinations of F, c, $\lambda(K)$ and α.

3.2 Model 1: Volume discounting targeted at high-usage customers

Our first proposition describes the optimal pricing model in region 1:

Region	**Restrictions on parameters**
No demand	$\frac{F}{\alpha} \leq \frac{1}{2}\left(\frac{c+\lambda(K)}{\alpha}\right)^2$
Region 1	$\frac{F}{\alpha} \geq \max\left\{\frac{1}{4}, \left(\frac{1}{2} - \frac{4}{9}\left[1 - \left(\frac{c+\lambda(K)}{\alpha}\right)\right]^2\right)\right\}$
Region 2	$\frac{1}{4} \geq \frac{F}{\alpha} \geq 4\left(\frac{c+\lambda(K)}{\alpha}\right)^2, \quad c+\lambda(K) \leq \frac{1}{4}$
Region 3	$\frac{1}{2}\left(\frac{c+\lambda(K)}{\alpha}\right)^2 \leq \frac{F}{\alpha} \leq \min\left\{4\left(\frac{c+\lambda(K)}{\alpha}\right)^2, \left(\frac{1}{2} - \frac{4}{9}\left[1 - \left(\frac{c+\lambda(K)}{\alpha}\right)\right]^2\right)\right\}$

Table 1: The analytical details of the regions mapped in Figure 1

Proposition 1 *In region 1 of the parameter space, when the cost of in-house deployment is high (relative to the service's business value of the service), the seller's optimal on-demand pricing schedule involves gradual volume discounting targeted at inducing high-usage customers to adopt. Specifically, the pricing function takes the form:*

$$P(q) = \begin{cases} \alpha\left[\left(\frac{[c+\lambda(K)]+\alpha}{2\alpha}\right)q - \frac{q^2}{4}\right] & \text{for } \frac{F}{\alpha} \geq \frac{1}{2} - \left(\frac{\alpha - c - \lambda(K)}{2\alpha}\right)^2 \\ \alpha\left[\left(1 - \sqrt{\frac{1}{2} - \frac{F}{\alpha}}\right)q - \frac{q^2}{4}\right] & \text{for } \frac{F}{\alpha} < \frac{1}{2} - \left(\frac{\alpha - c - \lambda(K)}{2\alpha}\right)^2 \end{cases} \tag{8}$$

The region of the parameter space under which Model 1 is optimal is illustrated in Figure 2. The pricing model in this region is "normal" in a sense: it involves no caps on demand, and a gradually increasing level of volume discounting. Notice that in the region above the dotted line, the pricing function does not depend on the fixed costs of in-house adoption. This is because the cost advantages of on-demand computing are sufficiently high (or alternatively, the costs of in-house deployment are sufficiently large) that the constraint placed on the seller by the presence of the in-house deployment alternative are no longer relevant, and pricing is according to standard unconstrained nonlinear pricing. Clearly, this will occur only if either the variable cost c of providing on-demand computing is low, or the seller has deployed a relatively high

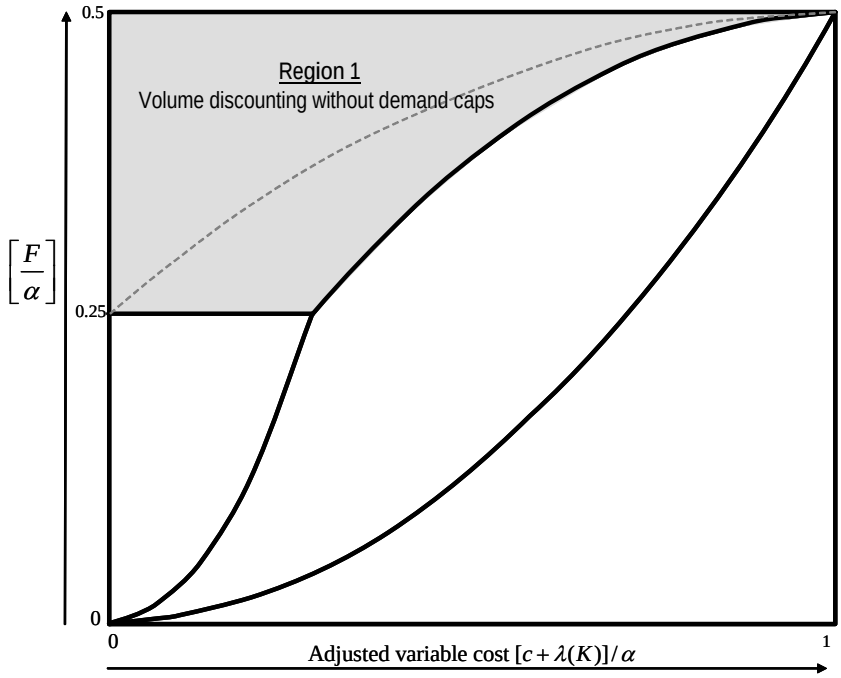

Figure 2: Illustrates the region of the parameter space under which Model 1 is applicable. The "in-house" deployment constraint does not bind above the dotted line, and pricing in the region above the line is according to the first line of the equation (8).

level of infrastructure (which would make $\lambda(K)$ relatively low), or both. However, the fraction of customer types who will adopt on-demand computing is concentrated among high-usage customers.

Below the dotted line, the alternative of in-house deployment begins to affect pricing. However, it is still optimal for the seller to price in a manner that targets all high-usage customers, although at a progressively lower per-unit price. Notice that in this case, the variable price

$$\frac{P(q)}{q} = \alpha \left[\left(1 - \sqrt{\frac{1}{2} - \frac{F}{\alpha}} \right) - \frac{q}{4} \right]$$

is strictly decreasing as F/α decreases because when developing in house becomes more attractive, the seller has to offer deeper discounts.

3.3 Model 2: Attracting a subset of high-volume users with free usage

Our next proposition describes the optimal pricing model in region 2:

Proposition 2 *In region 2 of the parameter space, when the business value of the service is high relative to the cost of in-house deployment, but the adjusted variable costs* $[c+\lambda(K)]$ *of providing on-demand computing are low, the seller's optimal on-demand pricing schedule involves gradual volume discounting at low levels of usage, free usage beyond a threshold, and often, a (relatively high) demand cap. Specifically, the pricing function takes the form:*

$$P(q) = \begin{cases} \alpha\left[\sqrt{\frac{F}{\alpha}}q - \frac{q^2}{4}\right] & \text{for } q < 2\sqrt{\frac{F}{\alpha}}; \\ F & \text{for } 2\sqrt{\frac{F}{\alpha}} \leq q \leq \min\left\{1, \frac{F}{c+\lambda(K)}\right\}. \end{cases} \tag{9}$$

The region of the parameter space under which Model 2 is appropriate, along with the general shape of the pricing function is illustrated in Figure 4. The pricing function therefore has upto three distinct regions. For low levels of usage, pricing is relatively standard, with a steadily increasing discount. However, at a threshold value of demand per customer, the seller stops increasing their total price, and offers free usage beyond that point. Notice that it is important for this segment of free usage to be targeted at the higher-usage customers – that is, those who have paid for their usage of the first $[2\sqrt{F/\alpha}]$ units of usage – rather than it being offered as "use k units for free up front and then start paying" schedule. This is because its purpose is to get the high-usage customers to find the on-demand computing alternative at least as attractive as their in-house alternative, while still being able to attract sufficient numbers of low-usage customers.

In the region above the dotted line of Figure 3 (on the left), the demand cap $(F/[c+\lambda(K)])$ from line 2 of equation (9) is non-binding – that is, it is higher than the maximum level of demand from any customer type. However, below this dotted line, this "demand cap" becomes binding. This is because in this region, the firm no longer finds it profitable to induce the

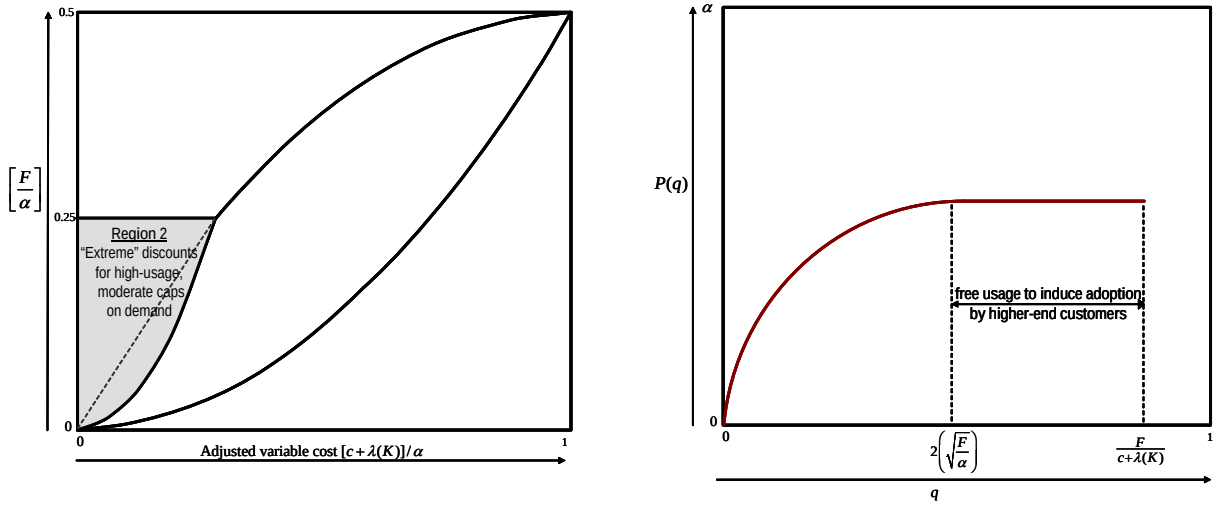

Figure 3: Illustrates the parameter region under which Model 2 applies, and the corresponding pricing function in this region. The demand cap illustrated in the pricing figure on the right is only present in the region of the parameter space below the dotted line .

highest-usage customers to adopt on-demand computing – their variable costs are too high to profitably offer usage beyond the usage level $(F/[c+\lambda(K)])$ while still charging a maximum total price F. The pricing strategy therefore transitions towards being focused on the "middle of the market", and away from the higher end. This is a trend that is accentuated further in Region 3, as described in our next proposition.

3.4 Model 3: Inducing limited adoption by low-usage customers

Our next proposition describes the optimal pricing model in region 3:

Proposition 3 *In region 3 of the parameter space, when the business value of the service is not very high relative to the cost of in-house deployment, and the adjusted variable costs* $[c+\lambda(K)]$ *of providing on-demand computing are high, the seller's optimal on-demand pricing schedule involves gradual volume discounting at low levels of usage, and a very restrictive demand cap.*

Specifically, the pricing function takes the form:

$$P(q) = \alpha \left[\left(\frac{[c + \lambda(K)] + \alpha\gamma}{2\alpha} \right) q - \frac{q^2}{4} \right] \text{ for } q < \frac{\gamma^2 - (\frac{c+\lambda}{\alpha})^2}{2\frac{(c+\lambda)}{\alpha}}, \tag{10}$$

where γ is a positive root of the equation

$$4\frac{(c+\lambda)}{\alpha}\widehat{\theta}(\gamma) - (\gamma + \frac{c+\lambda}{\alpha})^2 = 0; \tag{11}$$

$$\widehat{\theta}(\gamma) = \gamma + \frac{(c+\lambda)}{\alpha} - \frac{\sqrt{2}}{2}\sqrt{(\gamma + \frac{c+\lambda}{\alpha})^2 - 4\frac{F}{\alpha}}. \tag{12}$$

While we do not have a closed-form solution for the equation, it is not essential for us to describe the structure of the pricing function. We have also performed extensive numerical analysis to study the extent to which the demand cap varies, and find that indeed, it is quite restrictive, allowing participation by only a subset of customer types.

The pricing function and relevant region are illustrated in Figure 4. The intuition for this demand cap is quite straightforward – it is due to the fact that the adjusted variable costs are high relative to the cost of in-house fulfilment. As a consequence of these high costs, the seller cannot profitably sell its service to the higher-usage customers while still ensuring that its price is below their outside alternative (in-house development). However, it can sell profitably to those customers who cannot afford the in-house alternative at all – these are the lower usage customers. Clearly, without a demand cap, a price that is affordable to these customers would also be attractive to the higher usage customers, and consequently, in order to profit the most from their sales to these customers, the structure in (10) emerges.

Our final proposition defines the region of the parameter space for which on-demand computing is not profitable:

Proposition 4 *There is no demand for on-demand computing if and only if $\frac{F}{\alpha} < \frac{1}{2}(\frac{c+\lambda}{\alpha})^2$.*

In this region, the cost of providing on-demand computing is high enough relative to the cost of in-house deployment to render its provision unprofitable for any customer type.

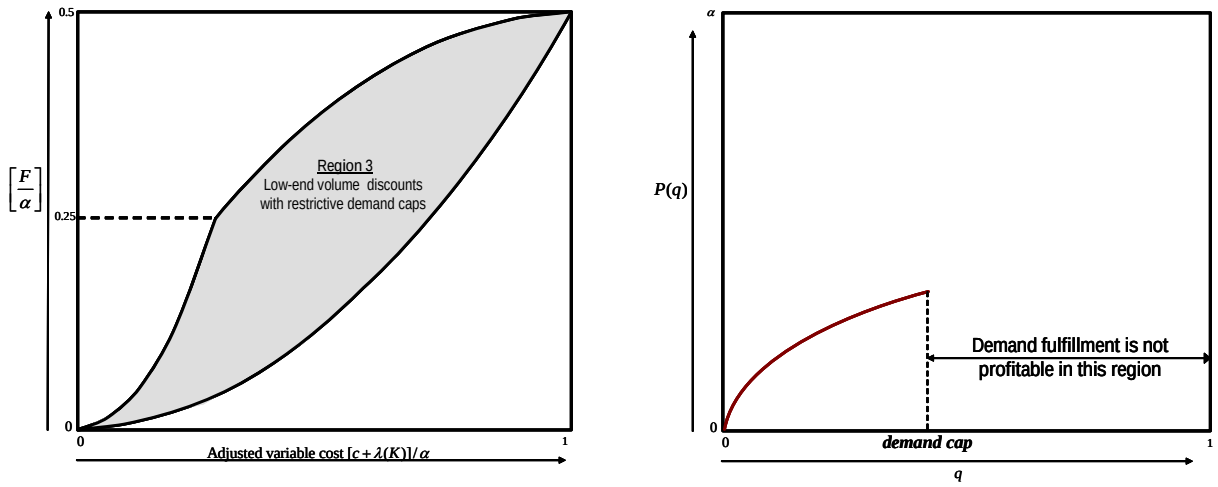

Figure 4: Illustrates the region of the parameter space under which Model 3 is applicable, and the corresponding optimal pricing function. The value of the demand cap increases as adjusted variable costs $[c + \lambda(K)]$ decrease, that is, as either variable costs fall or the limits on demand fulfilment imposed by infrastructure are relaxed.

4 Discussion, a possible path of transition and concluding remarks

As described in the brief discussion following the propositions, each of these three pricing models is optimal for a specific set of relative levels of average business value, variable cost, seller infrastructure, and fixed cost of in-house deployment. The sequence of pricing models that a seller will use, and correspondingly, the set of customer segments they target with their pricing strategy, depends on how these relative levels change over time.

Figure 5 summarizes a possible trajectory for the maturing of on-demand computing. In the years before technological advances in grid computing and Internet access made the model viable, it was clear that the variable costs of on-demand computing were too high to base a viable business around it, and the region in which it is not viable is reflective of this. Next, consider an industry in which the average business value α is relatively high. In the early stages of evolution of an industry's transition to on-demand computing, provider scale will be relatively low, and additionally, the variable costs of administering its provision are likely to be relatively

high, prior to any economies of scale that arise out of learning by doing. As a consequence, it seems likely that providers will be in region 3 at this stage. Our model prescribes that their optimal pricing strategy should be designed to attract lower-usage customers, by offering demand-capped pricing that draws in those companies who lack the scale to implement their own in-house solutions. In fact, there may be another reason for this approach early in the process: it gives providers the opportunity to gain expertise and "iron out the kinks" in their provision models without running the risk of a high-profile failure that could stifle a nascent industry.

As the variable costs of provision fall over time, and providers are sufficiently confident in the longer-term viability of the on-demand computing model to invest in sufficient scale to decrease $\lambda(K)$, the industry is likely to transition into region 2. During this phase, providers should design their pricing strategy to slowly expand the middle of the market, towards eventually fulfilling demand from higher-usage and larger corporate customers. Our model recommends that this expansion can be implemented most profitably by offering limited free-usage to customers whose individual demand exceeds a pre-specified level. This ensures that sellers can continue to profit from the lower-usage customers they have attracted early in the process, while being an economically viable alternative for these larger customers they are slowly drawing in.

Finally, as an increasing fraction of companies begin to transition to an on-demand corporate computing infrastructure, it is likely that the *fixed costs* of in-house deployment will *rise*, owing to the shrinking of the IT industry segments that support this kind of in-house provision. Indeed, a similar transition in cost structure occurred in the electricity provision industry in the early 20th century[4]. Prior to the widespread availability of centralized utility-based power,

[4]Our analogy has only to do with the similarity in cost structure, and unlike Carr (2005), we do not imply any similarity in the importance of IT and electricity to business. It is clear that the latter general-purpose technology, information technology, actually becomes progressively *more* important as shared infrastructures (like those supporting widespread on-demand computing) become available (Dhar and Sundararajan, 2005).

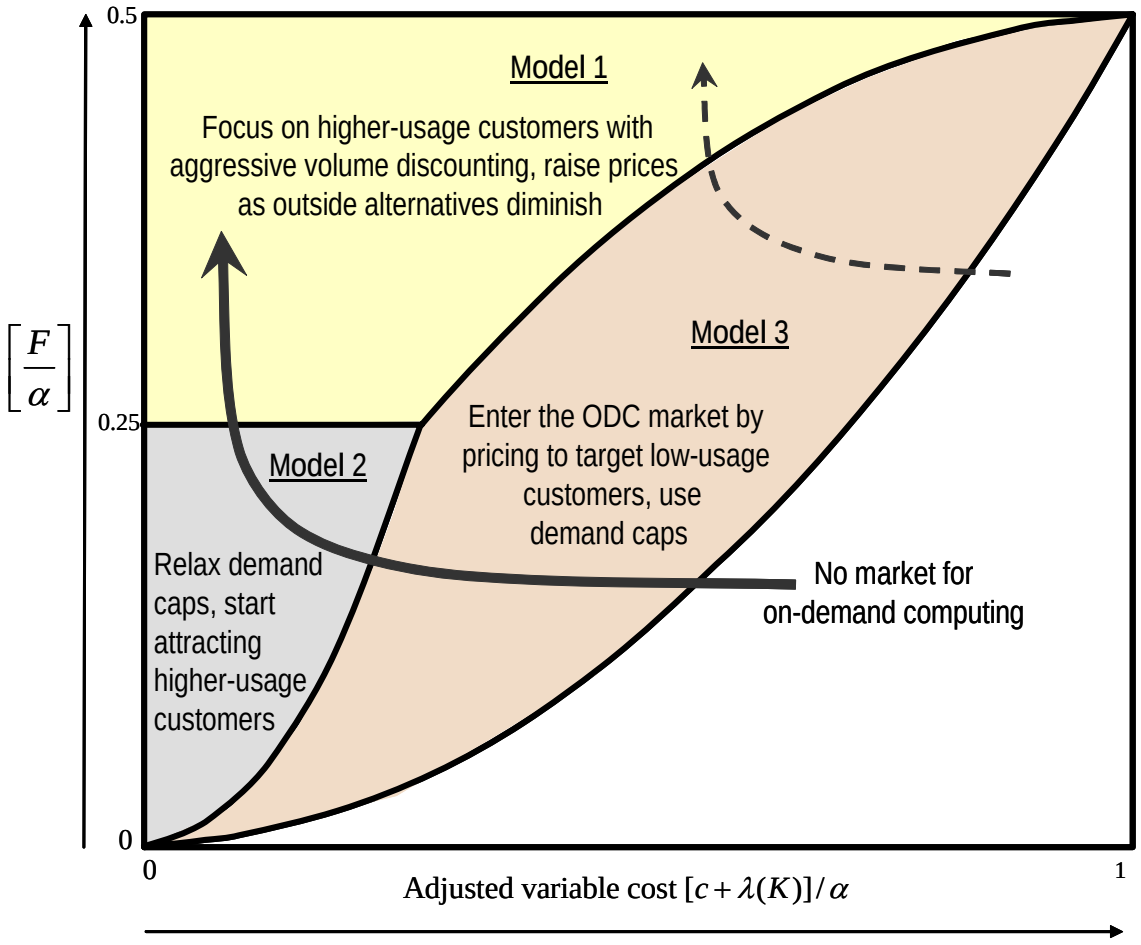

Figure 5: Summarizes a possible path of evolution to on-demand computing in an infrastructure industry with relatively high average business value, and the corresponding transition in pricing models and customer adoption that our model prescribes for this path. The solid line depicts a slower pace and transition through all three models, which is likely for IT infrastructure with higher average business value α relative to its costs and which is more tightly integrated with a customer's business, while the dotted line depicts a more rapid transition, more likely when pricing IT infrastructure for which willingness to pay is closer to its cost of provision.

many companies operated their own power plants. As the centralized power transmission and distribution infrastructure matured, and more and more larger companies switched to buying electricity on-demand (as the vast majority of them do today), the costs of setting up and running one's own dedicated power plant became prohibitively high. If the in-house IT provision industry evolves in a similar way (where the change in cost structure will be driven largely by the increasing cost of managing such an infrastructure in-house as they become progressively less common), the industry will transition into region 1, and pricing is likely to become more "normal": aimed at a broad segment of high-usage customers, with volume discounts and no artificial caps on demand.

In contrast, for an industry in which the average business value α of the service is relatively lower, the transition to region 3 may occur far more rapidly, since as indicated by our model, it is the ratios $[F/\alpha]$ and $[c + \lambda(K)]/\alpha$ that influence which region of the parameter space the firm is in. For example, for lower values of α, and a comparable cost structure, the transition is likely to be more rapid, and directly from region 1 to region 3. This leads to an interesting and intuitive conclusion: that on-demand company is likely to be widely prevalent earlier in those sectors of the IT infrastructure business which are less critically tied to a company's core value propositions. These are likely to be in commodity-like sectors such as web hosting, large scale processing, and data storage. Other forms of computing more closely tied to a company's business and customers are likely to remain as a mix – large-scale in-house deployment coexisting with on-demand alternatives for lower-end/mid-range customers – for much longer. Supply-chain management and CRM seem like good examples.

Our analysis suggests many directions for future research. The model we have articulated might apply equally well to a study of the optimal pricing of digital goods as a service; the analogue of F would be the price of the tangible (or unlimited usage) counterpart. Indeed, such usage-based business models are emerging in digital music and video. The prior literature (for

instance, Sundararajan, 2004) has characterized how these usage-based and unlimited usage options can coexist when offered by the same vendor. However, very little is known about the pricing strategy for an entrant whose business model is based on separating a digital good from its artifact and selling it as a service, when the level of F to be strategically varied by the incumbent. Since such entry threats are likely to become more common over time, this represents an interesting direction for further inquiry.

In order to focus our analysis on pricing structure, we ignore the possibility of peak-load pricing (for instance, like in Oren, Smith and Wilson, 1985) and of variability in the level of usage of individual customers. These two issues are clearly related. An extension of our model that accounts for these factors is a possible direction for future researchers, and an interesting one, given that such pricing is widely used for electricity industry, although such pricing is probably more important in the electricity industry, where adjusting "infrastructure" (the capacity to generate power) is significantly more difficult.. A limitation of our model is that it does not relate the fixed costs of in-house deployment to the scale of the customer explicitly. As we argue in Section 2, this may not be a significant assumption if providers can design and offer different pricing models for segments containing customers of differing scale. A generalization might associate in-house deployment scale endogenously with cost (perhaps by relating it to θ), although our analysis indicates that such a generalization is likely to be mathematically quite challenging.

To summarize, we have presented a new set of models of pricing on-demand corporate computing that explicitly take into account the relative business value of the infrastructure in question, the fixed costs of in-house deployment, the variable costs of provision and the scale of provider infrastructure. Our analysis shows that there are three distinct pricing models that may be optimal for on-demand computing, each of which is likely to characterize a specific stage in the transition. These pricing models may feature volume discounting, free usage for

high-demand customers, and demand caps, and the emergence of these features validates our approach of choosing to place no restrictions on the choice of pricing function, despite the associated analytical complexity of this approach. We are able to characterize which model is suitable based on the stage of evolution of an infrastructure industry towards the on-demand model, and suggest a likely path for an industry making the transition. As more vendors and their customers make this transition, further research on this increasingly important area is natural, and something we hope to contribute towards.

5 References

1. Afeche, P. and Mendelson H., 2004. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science* 50, 869-882.

2. Anand, K. and Aron R., 2003. Group buying on the Web: A comparison of price-discovery mechanisms. *Management Science* 49 (11), 1546-1562.

3. Armstrong, M., 1996. Multiproduct nonlinear pricing. *Econometrica* 64 (1), 51-75.

4. Bakos, Y. and Brynjolfsson, E., 1999. Bundling information goods: pricing, profits and efficiency. *Management Science* 45 (12), 1613-1630.

5. Bhargava, H. and Choudhary, V., 2001. Information goods and vertical differentiation. *Journal of Management Information Systems* 18, 89-106.

6. Bhargava, H. and S. Sundaresan, 2004. Computing as utility: managing availability, commitment and pricing through contingent bid auctions. *Journal of Management Information Systems* 21 (2), 201-227.

7. Carr, N., 2005. The End of Corporate Computing. *Sloan Management Review* 46 (3), 67-73.

8. Chen, P. and Wu, S., 2004. New IT Architecture: Implications and Market Structure. Workshop on Information Systems and Economics.

9. Clemons, E. and Kleindorfer, P., 1992. An economic analysis of interorganizational information technology. *Decision Support Systems* 8 (5), 431-446.

10. Clemons, E., Reddi, S. and Rowe, M., 1995. The impact of information technology on the organization of economic activity: the "move to the middle" hypothesis.

11. Dhar, V. and Sundararajan, A., 2005. Does IT matter in business education? Interviews with business school deans. Mimeo, New York University

12. Dewan, S. and Mendelson, H., 1990. User delay costs and internal pricing for a service facility. *Management Science* 36, 1502-1517.

13. Dewan, S. 1996. Pricing computer services under alternative control structures: tradeoffs and trends. Information Systems Research 7 (3), 301-307.

14. Fudenberg, D. and Tirole, J., 1991. Game Theory. MIT Press.

15. Geng, X., Stinchcombe, M. and Whinston, A., 2005. Bundling information goods of decreasing value. *Management Science* 51, 662-667.

16. Gurnani, H. and Karlapalem, K., 2001. Optimal pricing strategies for Internet-based software dissemination. *Journal of the Operational Research Society*, 52 (1), 64-70.

17. Huang K. and Sundararajan A., (2005). Pricing information technology: discontinuous and declining costs. Mimeo, New York University.

18. Jullien B., 2000. Participation constraints in adverse selection models. *Journal of Economic Theory* 93 (1), 1-47.

19. Konana, P. Gupta, A. and Whinston, A., 2000. Integrating user preferences and real-time workload in electronic commerce. *Information Systems Research* 11, 177-196.

20. Lewis T. and Sappington D., 1989. Countervailing incentives in agency problems. *Journal of Economic Theory* 49, 294–313.

21. Maggi G. and Rodriguez A., 1995. On countervailing incentives. *Journal of Economic Theory* 66, 238–263.

22. Mendelson, H., 1985. Pricing computer services: queuing effects. *Communications of the ACM* 28, 312-321.

23. Mendelson, H. and Whang, S., 1990. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Operations Research* 38, 870-883.

24. Mussa, M., and Rosen, S., 1978. Monopoly and product quality. *Journal of Economic Theory* 18, 301-317.

25. Nault, B., 1997. Quality differentiation and adoption costs: the case for interorganizational information systems pricing. *Annals of Operations Research* 71, 115-142.

26. Oren, S., Smith, S. and Wilson, R., 1985. Capacity pricing. *Econometrica* 53, 545-566.

27. Paleologo, G., 2004. A methodology for pricing utility computing services. *IBM Systems Journal* 43 (1), 20-31.

28. Snir, E. and Hitt, L., 2003. Costly bidding in online markets for IT services. *Management Science* 49 (11), 1504-1520.

29. Sundararajan, A., 2004a. Nonlinear pricing of information goods. *Management Science* 50, 1660-1673.

30. Sundararajan, A., 2004b. Managing digital piracy: pricing and protection. *Information Systems Research* 15, 287-304.

31. Weber, T., 2001. Mixed versioning of information goods under incomplete information. *Proceedings of the 22nd International Conference on Information Systems.*

A Appendix: A precise formulation of the pricing problem

Given a choice of pricing function $P(q)$, recall that the set of customers who adopt on-demand computing are those whose type θ satisfies:

$$\max_q[u(q,\theta) - P(q)] \geq 0. \tag{13}$$

$$\max_q[u(q,\theta) - P(q)] \geq \left(\max_q[u(q,\theta)]\right) - F. \tag{14}$$

denote the subset of θ values which satisfy (13) and (14) as $\Theta(P)$. The demand from a customer of type $\theta \in \Theta(P)$ will be $q(\theta, P)$, where:

$$q(\theta, P) \equiv \arg\max_q u(q,\theta) - P(q). \tag{15}$$

The total demand that the seller receives is therefore:

$$\int_{\theta\in\Theta(P)} q(\theta, P)d\theta, \tag{16}$$

and the seller's total profits are therefore:

$$\int_{\theta\in\Theta(P)} [P(q(\theta, P)) - cq(\theta, P)]d\theta. \tag{17}$$

The seller's problem is to choose the function $P(.)$ that maximizes (17) subject to the constraint

$$\int_{\theta\in\Theta(P)} q(\theta, P)d\theta \leq Q(K). \tag{18}$$

or, summarizing, the seller's problem is

$$\text{[P1]} \quad \max_{P(.)} \int_0^1 \left[P(q(\theta, P)) - cq(\theta)\right] d\theta \tag{19}$$

subject to

$$q(\theta, P) \in \arg\max_q \left[u(q, \theta) - P(q)\right], \tag{20}$$

$$\max_q \left[u(q, \theta) - P(q)\right] \geq \max\left[\max_q \left[u(q, \theta) - F\right], 0\right], \tag{21}$$

$$\int_0^1 q(\theta, P) d\theta \leq Q(K). \tag{22}$$

The first constraint specifies that each customer chooses the quantity that maximizes his gross surplus. The second constraint specifies that each buyer buys from the seller only if his surplus from buying on-demand computing is higher than that from his alternative options which are developing the service in-house or doing nothing (= 0). The last constraint specifies that the seller cannot sell more than the capacity constraint allowed by his infrastructure resources, K.

Rather than considering all possible pricing functions, the revelation principle ensures that the seller can restrict its attention to direct mechanisms—that is, usage-based contracts in which one specific quantity-price pair is designed for each customer, and in which it is rational and optimal for the customer to choose the quantity price pair that was designed for him or her[5]. We can thus rewrite [P1] as [P1a].

$$\text{[P1a]} \quad \max_{p(\theta), q(\theta)} \int_0^1 \left[p(\theta) - cq(\theta)\right] d\theta \tag{23}$$

subject to (24)

$$\text{[IC]:} \quad q(\theta) \in \arg\max_t \left[u(q(t), \theta) - p(t)\right], \tag{25}$$

$$\text{[PC]:} \quad \arg\max_t \left[u(q(t), \theta) - p(t)\right] \geq \max\left[\max_q \left[u(q, \theta) - F\right], 0\right], \tag{26}$$

$$\text{[CC]:} \quad \int_0^1 q(\theta) d\theta \leq Q(K). \tag{27}$$

Followed a standard transformation from the nonlinear pricing literature (Armstrong 1996), define the surplus function as

$$s(\theta) \equiv u(q(\theta), \theta) - p(\theta), \tag{28}$$

[5] This kind of formulation is standard in models of price screening—see, for instance, Armstrong (1996, §2). A good exposition of mechanism design, the revelation principle, and its applications to pricing can be found in Fudenberg and Tirole (1991, Chapter 7).

and [P1a] can be rewritten as problem [P2]:

$$\text{[P2]} \qquad \max_{q(\theta),s(\theta)} \int_0^1 \left[u(q(\theta),\theta) - s(\theta) - c \cdot q(\theta)\right] d\theta \tag{29}$$

$$\text{subject to} \tag{30}$$

$$\text{[IC1]:} \qquad q_\theta(\theta) > 0, \tag{31}$$

$$\text{[IC2]:} \qquad s_\theta(\theta) = u_\theta(q,\theta) = \alpha q, \tag{32}$$

$$\text{[PC]} \quad : \quad s(\theta) \geq \widehat{s}(\theta) \equiv \max\left[\frac{1}{2}\alpha\theta^2 - F, 0\right], \tag{33}$$

$$\text{[CC]} \quad : \quad \int_0^1 q(\theta) d\theta \leq Q(K). \tag{34}$$

Lemma 1 will show that [P2] is equivalent to the following formulation [P3], which is our final one.

$$\text{[P3]} \qquad \max_{q(\theta),s(\theta)} \int_0^1 \left[u(q(\theta),\theta) - s(\theta) - [c + \lambda(K)]q(\theta)\right] d\theta \tag{35}$$

$$\text{subject to} \tag{36}$$

$$\text{[IC1]:} \qquad q_\theta(\theta) > 0, \tag{37}$$

$$\text{[IC2]:} \qquad s_\theta(\theta) = u_\theta(q,\theta) = \alpha q, \tag{38}$$

$$\text{[PC]:} \qquad s(\theta) \geq \widehat{s}(\theta) \equiv \max\left[\frac{1}{2}\alpha\theta^2 - F, 0\right]. \tag{39}$$

For brevity later on, we have denoted the outside opportunity by $\widehat{s}(\theta)$. [P3] is an optimal control problem in which $q(\theta)$ is the control variable, $s(\theta)$ is the state variable, [IC2] is the equation of motion (dynamics), and which has two inequality constraints.

B Appendix: Proofs

Outline of the proof of Lemma 1

[P2] is called the isoperimetric problem in the optimal control literature (Seierstad and Sydsæster, 1987, Chapter 4). The Lagrangian or generalized Hamiltonian without the constraint [PC] for this problem is[6]

$$H^{P2} = u(q,\theta) - s(\theta) - c[q(\theta)] + \lambda[Q(K) - q(\theta)] + \mu[u_\theta(q,\theta)]. \tag{40}$$

Correspondingly, the Hamiltonian for [P3] without the constraint [PC] is

$$H^{P3} = u(q,\theta) - s(\theta) - (c + \lambda)q(\theta) + \mu u_\theta(q,\theta). \tag{41}$$

[6] We ignore the monotonicity constraing [IC1], later verifying for each case that our solution satisfies it.

In Huang and Sundararajan (2005), we show that for any value of $Q(K)$, there exists a unique solution for λ, the marginal benefit of an additional unit of capacity. We further show that the marginal revenue curve $\lambda(K)$ is continuous and strictly decreasing. As a result, under most general cost structures for $Q(K)$ (including ones that are discontinuous and in "blocks") that generate a relationship between infrastructure and the ability to fulfil a constrained level of maximum demand, there exists a unique optimal solution of λ, and consequently, the optimal solution of [P2] is equivalent to that of [P3].

The addition of the [PC] constraints (33) and (39) do not change the equivalence of H^{P2} and H^{P3}, since it involves the addition of the same term $[\mu^2(\theta)]\left[s(\theta) - \max\{\frac{1}{2}\theta^2 - F, 0\}\right]$ to both equations. While $\lambda(K)$ is still decreasing, it is no longer continuous, which adds some technical complexity, but does not change the idea above. A more detailed exposition is available on request.

Outline of the proofs of our propositions

We first state the necessary conditions of our problem in Lemma (2). Next, we solve the case when [PC] does not bind, in Lemma (3). We characterize the optimal solution of $(q(\theta), s(\theta))$ in Lemma (4), (5), and (6). We establish that $q^*(\theta)$ is a parallel shift of the solution to the standard nonlinear pricing problem (which we refer to subsequently as the *standard problem*), although $s^*(\theta)$ is somewhat different. Based on these lemmas, we can express the profit function as a function of just one variable γ, the extent of this parallel shift. We find that there are three cases that completely describe our problem, depending on which constraint is binding. We characterize the optimal contract sequentially in Lemma (7), (8), and (9), and these lemmas lead directly to our propositions.

Lemma 2 *Define $\underline{\theta}$ and $\overline{\theta}$ as the lowest and highest customer types who adopt on-demand computing. The necessary conditions for [P3] are:*

$$\begin{aligned}
\frac{\partial H^{P3}}{\partial q(\theta)} &= u_q(q^*,\theta) - (c+\lambda) + \mu^1(\theta)u_{q\theta}(q^*,\theta) = 0, \\
&\Rightarrow \quad q^*(\theta) = \theta + \mu^1(\theta) - \frac{c+\lambda}{\alpha}, && (42)\\
\mu^1_\theta(\theta) &= -\frac{\partial H^{P3}}{\partial s(\theta)} = 1 - \mu^2(\theta), && (43)\\
\textit{Transversality Conditions:} \quad & 0 = \mu^1(\underline{\theta})\left[s(\underline{\theta}) - \widehat{s}(\underline{\theta})\right], \ (\mu^1(\underline{\theta}) > 0 \Rightarrow s(\underline{\theta}) = \widehat{s}(\underline{\theta})), && (44)\\
0 &= \mu^1(\overline{\theta})\left[s(\overline{\theta}) - \widehat{s}(\overline{\theta})\right], \ (\mu^1(\overline{\theta}) > 0 \ \Rightarrow \ s(\overline{\theta}) = \widehat{s}(\overline{\theta})), && (45)\\
0 &= H^{P3}(q^*(\underline{\theta}), \mu^1(\underline{\theta}), \underline{\theta}), && (46)\\
0 &= H^{P3}(q^*(\overline{\theta}), \mu^1(\overline{\theta}), \overline{\theta}), && (47)\\
\textit{Kuhn-Tucker Condition:} \ 0 &= \mu^2(\theta)\cdot[s(\theta) - \widehat{s}(\theta)], \ (\mu^2(\theta) > 0 \ \Rightarrow \ s(\theta) = \widehat{s}(\theta)). && (48)
\end{aligned}$$

Proof. These conditions follow from the Maximum Principle, found in standard optimal control theory textbooks (Seierstad and Sydsæster, Chapter 5, for instance). Equations (42) and (43) indicate that the optimal quantity for each type will balance the sum of the marginal benefit from this type and the cost this type imposes indirectly on account of higher information rent to higher types. (44) to (47) are special in our problem since our boundary conditions on $\underline{\theta}$ and $\overline{\theta}$ are endogenously determined rather than fixed. The participation constraint adds a Kuhn-Tucker condition for this inequality constraint. ■

Lemma 3 *$(q^0(\theta), s^0(\theta))$ is the optimal solution if and only if*

$$\frac{(\alpha-\lambda-c)^2}{4\alpha} \geq \frac{1}{2}\alpha - F, \tag{49}$$

where

$$\begin{aligned} q^0(\theta) &= 2\theta - 1 - \frac{(c+\lambda)}{\alpha}, & (50)\\ s^0(\theta) &= \frac{\alpha}{4}\left(2\theta - 1 - \frac{c+\lambda}{\alpha}\right)^2. & (51) \end{aligned}$$

This also defines the optimal pricing strategy when the buyer does not have the option of in-house deployment (or the outside opportunity is zero for all buyers).

Proof. First, consider the solution to the problem in which $\widehat{s}(\theta) = 0$ for each θ. It is well known that $s(\underline{\theta}) = 0$ (Mussa and Rosen,1978). As a result, we have one transversality condition $\mu(1) = 0$. The necessary and sufficient conditions of (41) reduce to

$$\begin{aligned} \frac{\partial H^{P3}}{\partial q(\theta)} &= \alpha\left(\theta - q^0\right) - (c+\lambda) + \alpha\cdot\mu(\theta) = 0,\\ &\Rightarrow \quad q^0(\theta) = \theta + \mu(\theta) - \frac{c+\lambda}{\alpha}, & (52)\\ \mu_\theta(\theta) &= -\frac{\partial H^{P3\_0}}{\partial s(\theta)} = 1, & (53)\\ s(\underline{\theta}) &= 0, & (54)\\ \mu(1) &= 0. & (55) \end{aligned}$$

Consequently,

$$\mu(\theta) = \theta - 1, \tag{56}$$

and thus

$$q^0(\theta) = 2\theta - 1 - \frac{c+\lambda}{\alpha}. \tag{57}$$

Since $q^0(\underline{\theta}) = 0$, it follows from (44) that

$$\underline{\theta} = \frac{c+\lambda+\alpha}{2\alpha}. \tag{58}$$

The surplus function is thus

$$s^0(\theta) = 0 + \int_{\underline{\theta}}^{\theta} q^0(t)dt = \frac{\alpha}{4}\left(2\theta - 1 - \frac{c+\lambda}{\alpha}\right)^2. \tag{59}$$

Also notice that the condition in (49) is equivalent to $s^0(1) \geq \widehat{s}(1)$. Recall the definition of $\widehat{s}(\theta)$

$$\widehat{s}(\theta) = \frac{1}{2}\alpha\theta^2 - F, \tag{60}$$

and it therefore follows that:

$$s^0_\theta(\theta) - \widehat{s}_\theta(\theta) = \alpha\left[\theta - \left(1 + \frac{c+\lambda}{\alpha}\right)^2\right], \tag{61}$$

which is strictly negative since $\theta \leq 1$. From (61), $s^0(1) \geq \widehat{s}(1)$ implies $s^0(\theta) \geq \widehat{s}(\theta)$ for all $\theta \in [0,1]$, and the solution to the standard problem is feasible for all θ, which proves the necessary part. The sufficient part follows from the fact that if $s^0(1) < \widehat{s}(1)$, [PC] is violated for some $\theta \leq 1$. ■

Lemma 4 *Define* $\widehat{\theta} \equiv \min\{\theta : s^*(\theta) = \widehat{s}(\theta) > 0\}$. *If* $\hat{\theta}$ *exists, then the optimal quantity and price schedule takes the following form:*

$$q^*(\theta) = \begin{cases} 2\theta - \gamma - \frac{(c+\lambda)}{\alpha}, & \theta \in [\underline{\theta}, \widehat{\theta}) \\ \theta, & \theta \in [\widehat{\theta}, \overline{\theta}], \end{cases} \tag{62}$$

or the optimal allocation to each type is either a parallel upwards shift of the allocation under the standard problem (with $\gamma < 1$ *representing the extent of this shift), or an allocation that induces usage at the maximum possible level. Also,*

$$p^*(\theta) = \begin{cases} \frac{\alpha}{4}\left(3\frac{c+\lambda}{\alpha} + 3\gamma - 2\theta\right)q(\theta), & \theta \in [\underline{\theta}, \widehat{\theta}) \\ F, & \theta \in [\widehat{\theta}, \overline{\theta}], \end{cases} \tag{63}$$

or the optimal price to each type is either the price under the standard problem adjusted for the demand shift, or a price that replicates the cost of in-house deployment.

Proof. When $\theta < \widehat{\theta}$, $s^*(\theta) > \widehat{s}(\theta)$ by definition. and hence $\mu^2(\theta) = 0$ from (48). From (43), this implies that $\mu^1(\theta)$ must take the form $[\theta - \gamma]$, where γ is an arbitrary constant (we choose $[\theta - \gamma]$ rather than $[\theta + \gamma]$ since this makes γ positive later on). Substituting into (42) yields the first line of (62), and the price schedule in the first line of (63) follows accordingly based on the fact that

$$s^*(\theta) = s^*(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} \alpha q(t)dt, \tag{64}$$

$$p^*(\theta) = U(q^*(\theta), \theta) - s^*(\theta). \tag{65}$$

When $\theta > \widehat{\theta}$, the seller cannot set $s^*(\theta) > \widehat{s}(\theta)$ since this violates [PC]. As a consequence, the seller either chooses $s^*(\theta) = \widehat{s}(\theta)$, which can only be implemented by the quantity and price schedule in the second line of (62) and (63), or excludes the customer type θ, in which case $\theta > \overline{\theta}$. The result follows. ■

Lemma 5 *(1)* $s^*(\underline{\theta}) = 0$, *(2)* $q^*(\underline{\theta}) = 0$, *and (3)* $\underline{\theta} = \frac{1}{2}\left[\gamma + \frac{(c+\lambda)}{\alpha}\right]$.

Proof. (1) When $\underline{\theta} = 0$, it is not optimal to set $s^*(\underline{\theta}) > 0$. Next, suppose for some $\underline{\theta} > 0$, the seller chooses $s^*(\underline{\theta}) > 0$. If this is the case, it follows that $q^*(\underline{\theta}) > 0$, and since the seller sets a non-zero price,

$$u(q^*(\underline{\theta}), \underline{\theta}) - (c + \lambda)q^*(\underline{\theta}) > 0, \tag{66}$$

which, using the first line of (62), in turn implies

$$\alpha\left(\underline{\theta} - \frac{1}{2}q^*(\underline{\theta}) - \frac{(c+\lambda)}{\alpha}\right) q^*(\underline{\theta}) > 0, \tag{67}$$

or

$$\gamma - \frac{(c+\lambda)}{\alpha} > 0. \tag{68}$$

Consequently, there exists an $\varepsilon > 0$ such that $\underline{\theta} - \varepsilon > 0$ and $q^*(\underline{\theta} - \varepsilon) \in (0, q^*(\underline{\theta}))$. If the seller chooses $s^*(\underline{\theta} - \varepsilon) = 0$, the profit from serving type $\underline{\theta} - \varepsilon$ is

$$u(q^*(\underline{\theta} - \varepsilon), \underline{\theta} - \varepsilon) - s^*(\underline{\theta} - \varepsilon) - (c + \lambda)q^*(\underline{\theta} - \varepsilon), \tag{69}$$

which simplifies to

$$= \frac{\alpha}{2}\left(\gamma - \varepsilon - \frac{(c+\lambda)}{\alpha}\right) q^*(\underline{\theta} - \varepsilon), \tag{70}$$

which is strictly positive for a small enough ε. Introducing this contract does not violate [IC2], since it can be verified that

$$u(q^*(\underline{\theta} - \varepsilon), \underline{\theta}) - p^*(\underline{\theta} - \varepsilon) = \alpha\varepsilon q^*(\underline{\theta} - \varepsilon), \tag{71}$$

and the RHS of (71) can be made smaller than $s^*(\underline{\theta})$ by choosing an arbitrarily small ε. Therefore, including the customer types $[\underline{\theta} - \varepsilon, \underline{\theta}]$ improves the seller's profits, a contradiction, and the result follows.

(2) Given $s^*(\underline{\theta}) = 0$, by (46), we have

$$u(q, \underline{\theta}) - (c + \lambda) \cdot q(\underline{\theta}) + \mu^1(\underline{\theta})u_\theta(q, \underline{\theta}) = 0, \tag{72}$$

where the last term in H^{P3} is dropped because of (48). After substituting the expression from (62) and simplifying, (72) is equivalent to

$$\frac{1}{2}q(\underline{\theta})\left[\frac{1}{2}q(\underline{\theta}) - \underline{\theta}\right] = 0, \tag{73}$$

and the result follows since $q(\theta) \leq \theta$ under any pricing schedule.
(3) Follows immediately from the fact that $q(\underline{\theta}) = 0$. ■

Lemma 6 $\overline{\theta} = \min\left[q^{*-1}(\frac{F}{c+\lambda}), 1\right]$, *where* $q^{*-1}(\theta)$ *is the inverse of* $q^*(\theta)$

Proof. Given any $q(\theta)$, the highest level of profit feasible from type θ is

$$u(q(\theta), \theta) - [c + \lambda(K)]q(\theta) - \hat{u}(\theta), \tag{74}$$

which simplifies to

$$-\frac{1}{2}\alpha[\theta - q(\theta)]^2 + [F - [c + \lambda(K)]q(\theta)]. \tag{75}$$

If $q(\theta) > F/[c + \lambda(K)]$, (75) is strictly negative, which proves that $\overline{\theta} \leq q^{*-1}(\frac{F}{c+\lambda})$. When $\overline{\theta} = q^{*-1}(\frac{F}{c+\lambda})$, we can show(later) that $s^*(\overline{\theta}) = \hat{s}(\overline{\theta})$ except in the standard case. As a consequence, by lemma (4), the profit is zero at $\overline{\theta}$. By assumption, the marginal customer is served and thus $\overline{\theta} = q^{*-1}(\frac{F}{c+\lambda})$ when $\overline{\theta} \leq 1$. Otherwise, $\overline{\theta} = 1$. ■

We can now rewrite the profit function in terms of only one argument, γ. We are going to show that, with the exception of one case, separately analyzed in Proposition 3, there are two regions of the profit function and the optimal value of γ is at the boundary of those two regions. Define $q^*(\theta, \gamma)$ as the demand from a customer of type θ and $s^*(\theta, \gamma)$ as the consumer surplus for type θ when the shift γ is chosen in (62). Correspondingly, following Lemma 4, define $\hat{\theta}(\gamma) = \min\{\theta : s^*(\theta, \gamma) = s(\theta)\}$. Define the following critical values of γ:

$$\gamma_1 = \min\{\gamma : s^*(\theta, \gamma) = \hat{s}(\theta)\} \tag{76}$$

where γ_1 is the lowest such value for some $\theta \in [0, 1]$, and

$$\gamma_2 = \gamma : s^*(1, \gamma) = \hat{s}(1). \tag{77}$$

The value of γ_1 in (76) may not always exist. In the cases that follow, the variable γ_0 will take the value either γ_1 or γ_2 and we will show they are the solutions. Our problem is therefore to maximize $\pi(\gamma)$ with respect to γ, and has the following cases, which we label based on their eventual solutions:

Case 1: $\frac{(\alpha-\lambda-c)^2}{4\alpha^2} \geq \frac{1}{2} - \frac{F}{\alpha}$ (standard pricing): The solution is stated in Lemma (3).

Case 2: $\frac{(\alpha-\lambda-c)^2}{4\alpha^2} < \frac{1}{2} - \frac{F}{\alpha}$ and $\frac{F}{\alpha} \geq \min(\frac{c+\lambda}{\alpha}, \left(\frac{1}{2} - \frac{4}{9}\left[1 - \frac{c + \lambda(K)}{\alpha}\right]^2\right))$[7] (non-standard with possible free usage, and no demand cap): The profit function has two regions:

Region 1, when $\gamma \geq \gamma_0$,

$$\pi(\gamma) = \int_{\underline{\theta}}^{\widehat{\theta}(\gamma)} [u(q^*(\theta, \gamma), \theta) - [c + \lambda]q^*(\theta, \gamma) - s^*(\theta)]\, d\theta + \int_{\widehat{\theta}(\gamma)}^{1} [u(\theta, \theta) - [c + \lambda]\theta - \widehat{s}(\theta)]\, d\theta, \tag{78}$$

[7] With these constraints, we can verify that the solutions satisfy all constraints. $\frac{F}{\alpha} \geq \frac{c+\lambda}{\alpha}$ comes from $\overline{\theta} \leq 1$ within region 2 and $\frac{F}{\alpha} \geq \frac{1}{2} - \frac{4}{9}\left[1 - \frac{c + \lambda(K)}{\alpha}\right]^2$ comes from $p(q^*(1)) - [c + \lambda(K)]\, q^*(1) \geq 0$ in region 1.

Region 2, when $\gamma < \gamma_0$,

$$\pi(\gamma) = \int_{\underline{\theta}}^{\min\left[1,(\gamma+\frac{c+\lambda}{\alpha}+\frac{F}{c+\lambda})/2\right]} [u(q^*(\theta,\gamma),\theta) - [c+\lambda]q^*(\theta,\gamma) - s^*(\theta,\gamma)]\, d\theta. \tag{79}$$

The upper bound comes from lemma 4 and 6. When $\frac{F}{\alpha} \leq \frac{1}{4}$[8] , $\gamma_0 = \gamma_1$ (**case 2A**), and when $\frac{F}{\alpha} > \frac{1}{4}$, $\gamma_0 = \gamma_2$ (**case 2B**). The intuition is that when γ is smaller, $q^*(\theta,\gamma)$ is larger and $s^*(\theta,\gamma)$ is greater than $\widehat{s}(\theta)$ for each θ. As a result, there exists a threshold value of γ, whose value is γ_1, such that the objective function changes its shape. At the same time, the constraint $\widehat{\theta}(\gamma) \leq 1$ leads to Case 2B, with the threshold determined by $\widehat{\theta}(\gamma) = 1$. We will show that the profit function is increasing in region 1 and decreasing in region 2. As a consequence, we will have a corner solution at $\gamma^* = \gamma_0$.

Case 3: $\frac{F}{\alpha} < \min(\frac{c+\lambda}{\alpha}, \left(\frac{1}{2} - \frac{4}{9}\left[1 - \frac{c+\lambda(K)}{\alpha}\right]^2\right))$, (non-standard pricing with a demand cap): Again, we have two subcases:

- **Case 3A:** When $\frac{2(c+\lambda)}{\alpha} \leq \sqrt{\frac{F}{\alpha}}$, the profit function has two regions, analogous to those in case 2. For $\gamma \geq \gamma_0$,

$$\pi(\gamma) = \int_{\underline{\theta}}^{\widehat{\theta}(\gamma)} [u(q^*(\theta,\gamma),\theta) - [c+\lambda]q^*(\theta,\gamma) - s^*(\theta)]\, d\theta + \int_{\widehat{\theta}(\gamma)}^{\frac{F}{c+\lambda}} [u(\theta,\theta) - [c+\lambda]\theta - \widehat{s}(\theta)]\, d\theta, \tag{80}$$

  and for $\gamma < \gamma_0$,

$$\pi(\gamma) = \int_{\underline{\theta}}^{\min\left[1,(\gamma+\frac{c+\lambda}{\alpha}+\frac{F}{c+\lambda})/2\right]} [u(q^*(\theta,\gamma),\theta) - [c+\lambda]q^*(\theta,\gamma) - s^*(\theta,\gamma)]\, d\theta, \tag{81}$$

  where $\gamma_0 = \gamma_1$.

- **Case 3B:** When $\frac{2(c+\lambda)}{\alpha} > \sqrt{\frac{F}{\alpha}}$, the profit function has three regions, and its discussion is deferred to the proof of Proposition 3.

These cases are mapped out onto our parameter space in Figure 6 to make the connection with the Propositions clearer.

We now proceed to show that:

- Case (2A): $\widehat{\theta}(\gamma) < 1$, $\gamma_1 < 1$

[8] $\frac{F}{\alpha} \leq \frac{1}{4}$ comes from $\widehat{\theta}(= 2\sqrt{\frac{F}{\alpha}}) \leq 1$ in region 2.

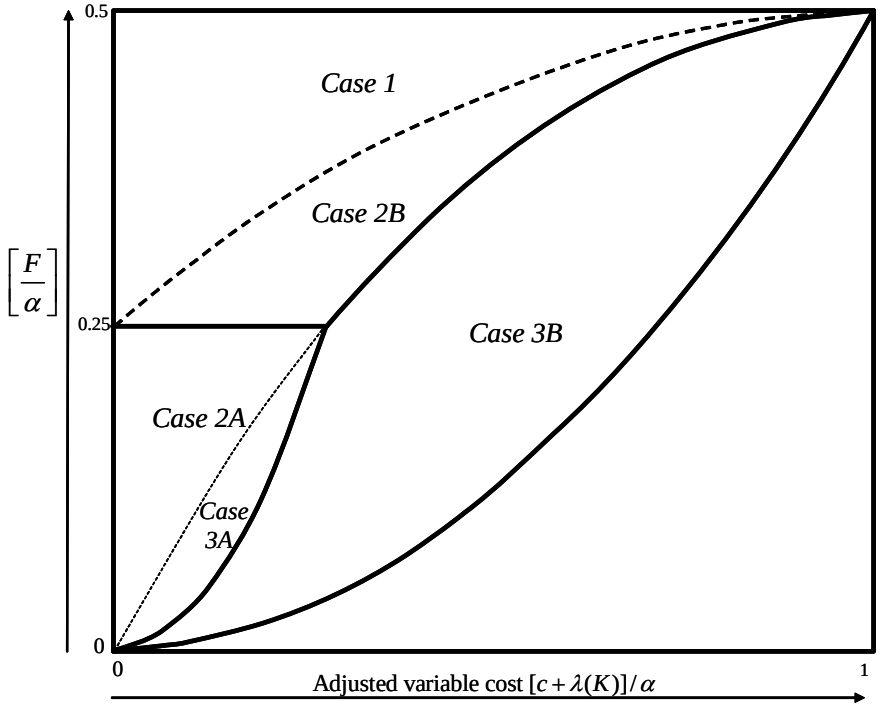

Figure 6: Mapping of our cases in the Appendix to the regions of the parameter space that each case describes the solution to.

- Case (2B): $\widehat{\theta}(\gamma) > 1$, $\gamma_2 < \gamma_1 < 1$
- Case (3A), $\widehat{\theta}(\gamma) < F/(c+\lambda), \gamma_1 < 1$.
- Cases (2A), (2B) and (3A): The profit function is increasing for all $\gamma < \gamma_0$, and the profit function is decreasing for all $\gamma > \gamma_0$.

We summarize these proofs in the following three lemmas.

Lemma 7 *When* $\frac{(\alpha-\lambda-c)^2}{4\alpha} < \frac{1}{2}\alpha - F$ *and* $\gamma < \gamma_0$, $\pi(\gamma)$ *is increasing in* γ.

Proof. Independent of whether $\gamma_0 = \gamma_1$ or γ_2, when $\overline{\theta} = 1$, the profit function in (79) simplifies to

$$\pi(\gamma) = \frac{\alpha}{12}\left(2\gamma - 1 - \frac{c+\lambda}{\alpha}\right)\left[\gamma + \frac{(c+\lambda)}{\alpha} - 2\right]^2. \tag{82}$$

Differentiating with respect to γ and simplifying yields

$$\pi_\gamma(\gamma) = \left(\frac{\alpha}{2}\right)(1-\gamma)\left(2 - \gamma - \frac{(c+\lambda)}{\alpha}\right). \tag{83}$$

the second term is positive because $\gamma \leq 1$ is a standard result (Jullien 2000) and the third term in parenthesis on the RHS of (83) is positive because $q^*(1,\gamma) > 0$.

A similar computation of the profit function in (79) when $\overline{\theta} = (\gamma + \frac{c+\lambda}{\alpha} + \frac{F}{c+\lambda})/2$ yields

$$\pi_\gamma(\gamma) = \frac{\alpha F^2}{8\left(c+\lambda\right)^2}, \tag{84}$$

which is strictly positive. The result follows. ■

Lemma 8 *When* $\frac{(\alpha-\lambda-c)^2}{4\alpha} < \frac{1}{2}\alpha - F$ *and* $\gamma > \gamma_0$, $\pi(\gamma)$ *is decreasing in* γ.

Proof. We first identify $\widehat{\theta}(\gamma)$, the marginal type whose participation constraint is binding, given γ. This is the value of θ which solves $s^*(\theta,\gamma) = \widehat{s}(\theta)$, and this equation reduces to:

$$\frac{\alpha}{4}\left(2\widehat{\theta} - \gamma - \frac{(c+\lambda)}{\alpha}\right)^2 = \frac{1}{2}\alpha\widehat{\theta}^2 - F, \tag{85}$$

which yields:

$$\widehat{\theta}(\gamma) = \gamma + \frac{(c+\lambda)}{\alpha} - \frac{\sqrt{2}}{2}\sqrt{(\gamma + \frac{c+\lambda}{\alpha})^2 - 4\frac{F}{\alpha}}. \tag{86}$$

Next, we compute the following partial derivatives.

$$\frac{\partial\widehat{\theta}(\gamma)}{\partial\gamma} = \frac{\underline{\theta} - \widehat{\theta}(\gamma)}{[\gamma + \frac{(c+\lambda)}{\alpha} - \widehat{\theta}(\gamma)]} < 0. \tag{87}$$

$$\frac{\partial\underline{\theta}}{\partial\gamma} = \frac{1}{2} > 0. \tag{88}$$

$$\frac{\partial}{\partial\gamma}\left[u(q^*(\theta,\gamma),\theta) - (c+\lambda)q^*(\theta,\gamma) - s^*(\theta)\right] = 2\theta\alpha - \frac{(c+\lambda)}{2} - \frac{3}{2}\alpha\gamma. \tag{89}$$

Differentiating (78) with respect to γ yields:

$$\begin{aligned}\pi_\gamma(\gamma) \;=\; & \int_{\underline{\theta}}^{\widehat{\theta}}\left(2\theta\alpha - \frac{[c+\lambda]}{2} - \frac{3}{2}\alpha\gamma\right)d\theta + \left[\frac{3\alpha}{4}\gamma - \frac{\alpha}{2}\widehat{\theta}(\gamma) - \frac{[c+\lambda]}{4}\right][2\widehat{\theta}(\gamma) - \gamma - \frac{[c+\lambda]}{\alpha}] \\ & -\frac{\partial\widehat{\theta}(\gamma)}{\partial\gamma}\left[\frac{\alpha}{2}[\widehat{\theta}(\gamma)]^2 - \frac{\alpha}{4}\left(2\widehat{\theta}(\gamma) - \gamma - \frac{[c+\lambda]}{\alpha}\right)^2 - [c+\lambda]\widehat{\theta}(\gamma)\right],\end{aligned}$$

which, using (87-89) simplifies to:

$$\pi_\gamma(\gamma) = \frac{\alpha}{2}\left(\underline{\theta} - \widehat{\theta}\right)\left[\gamma + \frac{(c+\lambda)}{\alpha} - \widehat{\theta}\right],$$

which is strictly negative because the last term is positive by (86). A similar computation for Case 3A, which is omitted, yields the required result. ■

Lemma 9 *The value of γ that maximizes $\pi(\gamma)$ is, for each of the cases (2A), (2B), and (3A) respectively:*

$$\textit{(2A):} \qquad \gamma^* = \gamma_1 = 2\sqrt{\frac{F}{\alpha}} - \frac{c+\lambda}{\alpha}. \tag{90}$$

$$\textit{(2B):} \qquad \gamma^* = \gamma_2 = 2 - \frac{c+\lambda}{\alpha} - \sqrt{2 - \frac{4F}{\alpha}}. \tag{91}$$

$$\textit{(3A):} \qquad \gamma^* = \gamma_1 = 2\sqrt{\frac{F}{\alpha}} - \frac{c+\lambda}{\alpha}. \tag{92}$$

Proof. We need to show that under the conditions in each case, $\widehat{\theta}(\gamma^*) \leq 1$ and $\gamma^* < 1$. We first derive γ_1 and γ_2 from their definitions.

$$\gamma_1 = 2\sqrt{\frac{F}{\alpha}} - \frac{c+\lambda}{\alpha}. \tag{93}$$

$$\gamma_2 = 2 - \frac{(c+\lambda)}{\alpha} - \sqrt{2 - \frac{4F}{\alpha}}. \tag{94}$$

From (87), $\widehat{\theta}_\gamma(\gamma) < 0$ and therefore, $\widehat{\theta}(\gamma)$ is maximized at $\gamma = \gamma^*$.

Case 2A: $\frac{F}{\alpha} \leq \frac{1}{4}$ implies that $2\sqrt{\frac{F}{\alpha}} \leq 1$.

$$\widehat{\theta}(\gamma_1) = \gamma_1 + \frac{c+\lambda}{\alpha} = 2\sqrt{\frac{F}{\alpha}} \leq 1; \tag{95}$$

and

$$\gamma_1 + \frac{c+\lambda}{\alpha} \leq 1 \Rightarrow \gamma_1 \leq 1. \tag{96}$$

Case 2B: $\frac{F}{\alpha} > \frac{1}{4}$, $\widehat{\theta}(\gamma_2) = 1$ by definition. For γ_2, we know that $s^*(1, \gamma_2) = \widehat{s}(1)$ by definition. From (91), it follows that

$$\frac{1}{2}\alpha - F = \frac{\alpha}{4}\left(2 - \gamma_2 - \frac{(c+\lambda)}{\alpha}\right)^2, \tag{97}$$

which in conjunction with the condition $\frac{(\alpha-\lambda-c)^2}{4\alpha} < \frac{1}{2}\alpha - F$ implies that $\gamma_2 < 1$.

Case 3A: $\frac{2(c+\lambda)}{\alpha} \leq \sqrt{\frac{F}{\alpha}}$ implies that

$$\widehat{\theta}(\gamma_1) = 2\sqrt{\frac{F}{\alpha}} \leq \frac{F}{(c+\lambda)}. \tag{98}$$

The proof of $\gamma^* = \gamma_1 \leq 1$ is the same as that in (2A).

■

Proof of Proposition 1

This proposition corresponds to the regions of the parameter space covered by Case 1 and Case 2B. The optimal schedule for case 1 follows directly from Lemma (3). By definition, the optimal pricing schedule is simply

$$p^0(\theta) = \alpha(\theta q^0(\theta) - \frac{q^0(\theta)^2}{2}) - s^0(\theta). \tag{99}$$

Using the expressions for $q^0(\theta)$ and $s^0(\theta)$ from (50) and (51) in (99) enables us to derive the pricing function as a function of demand:

$$P^*(q) = \frac{1}{4}\alpha\left(2\frac{c+\lambda}{\alpha} + 2 - q\right)q. \tag{100}$$

The optimal schedule for case 2B follows from Lemmas (4), (7), (8), and (9). Substituting the optimal value γ^* into the expressions in (4) yields

$$\begin{aligned} q^*(\theta) &= 2\theta - 2 + \sqrt{2 - 4\frac{F}{\alpha}}; & (101)\\ s^*(\theta) &= \frac{1}{4}\alpha\left(2\theta - 2 + \sqrt{2 - 4\frac{F}{\alpha}}\right)^2; & (102)\\ p^*(\theta) &= \frac{3}{4}\alpha\left(2 - \sqrt{2 - \frac{4F}{\alpha}} - \frac{2\theta}{3}\right)q^*(\theta). & (103) \end{aligned}$$

Rearranging these expressions yields the pricing function as a function of demand:

$$P^*(q) = \frac{1}{4}\alpha\left[4 - 2\sqrt{2 - \frac{4F}{\alpha}} - q\right]q, \tag{104}$$

which completes the proof of the Proposition.

Proof of Proposition 2

The region of the parameter space that this Proposition corresponds to is covered by our cases 2A and 3A, as illustrated in Figure 6. In case 2A, from lemma 9,

$$\gamma^* = 2\sqrt{\frac{F}{\alpha}} - \frac{c+\lambda}{\alpha}, \tag{105}$$

by (86), the lowest customer type that receives free usage is:

$$\widehat{\theta} = 2\sqrt{\frac{F}{\alpha}} \tag{106}$$

and $\overline{\theta} = 1$. The optimal schedule then follows from Lemma (4):

$$q^*(\theta) = \begin{cases} 2(\theta - \sqrt{\frac{F}{\alpha}}), & \theta \in [\sqrt{\frac{F}{\alpha}}, 2\sqrt{\frac{F}{\alpha}}) \\ \theta, & \theta \in [2\sqrt{\frac{F}{\alpha}}, \overline{\theta}] \end{cases} . \tag{107}$$

$$s^*(\theta) = \begin{cases} \alpha \left[\theta - \sqrt{\frac{F}{\alpha}}\right]^2, & \theta \in [\sqrt{\frac{F}{\alpha}}, 2\sqrt{\frac{F}{\alpha}}) \\ \frac{1}{2}\alpha\theta^2 - F, & \theta \in [2\sqrt{\frac{F}{\alpha}}, \overline{\theta}] \end{cases} . \tag{108}$$

and therefore:

$$p^*(\theta) = \begin{cases} \alpha[3\sqrt{\frac{F}{\alpha}} - \theta][\theta - \sqrt{\frac{F}{\alpha}}], & \theta \in [\sqrt{\frac{F}{\alpha}}, 2\sqrt{\frac{F}{\alpha}}) \\ F, & \theta \in [2\sqrt{\frac{F}{\alpha}}, \overline{\theta}] \end{cases} . \tag{109}$$

Rearranging these expressions yields the pricing function as a function of demand:

$$P^*(q) = \frac{1}{4}\alpha \left(4\sqrt{\frac{F}{\alpha}} - q\right) \cdot q. \tag{110}$$

An identical sequence of steps yields the same solution for case 3A, with the exception that $\overline{\theta} = F/(c + \lambda)$ rather than 1.

Proof of Proposition 3

This proposition corresponds to case 3B. Define

$$\gamma_3 = \gamma : \widehat{\theta}(\gamma) = \frac{F}{(c + \lambda)}. \tag{111}$$

In contrast with cases 2A, 2B and 3A, the profit function may have up to three regions:

Region 1, when $\gamma \leq \gamma_1$, $s^*(\theta, \gamma) > \widehat{s}(\theta)$ for all θ, and the profit function is increasing as we showed in Lemma(7).

Region 2, when $\gamma_1 < \gamma \leq \min[\gamma_3, 1]$,then the demand cap is set such that all customers of type $\theta > \widehat{\theta}$ are excluded. We will show that the profit function has an interior maximum in this case.

Region 3, when $\gamma > \min\{1, \gamma_3\}$, $s^*(\widehat{\theta}) = \widehat{s}(\widehat{\theta})$, the exclusion happens at $\overline{\theta} > \widehat{\theta}$. The profit function is decreasing as we showed in Lemma(8). We may not have this case for some parameter values, but that does not matter, since the solution is never in this region.

Now, in Region 2, the profit function for a given γ is

$$\pi(\gamma) = \int_{\underline{\theta}}^{\widehat{\theta}(\gamma)} [u(q^*(\theta, \gamma), \theta) - [c + \lambda]q^*(\theta, \gamma) - s^*(\theta, \gamma)] \, d\theta. \tag{112}$$

Substituting the expression for quantity from Lemma (4), and then differentiating both sides of (112) with respect to γ yields:

$$\pi_\gamma(\gamma) = \frac{\left(4\widehat{\theta}c + 4\widehat{\theta}\lambda - (\gamma + \frac{c+\lambda}{\alpha})^2\alpha\right)\left(\widehat{\theta} - \underline{\theta}\right)}{4\left(\gamma + \frac{c+\lambda}{\alpha} - \widehat{\theta}\right)}. \tag{113}$$

Since $\widehat{\theta} > \underline{\theta}$ and using (80), the sign of the RHS of (113) depends only on the sign of

$$f(\gamma) \equiv 4\frac{(c+\lambda)}{\alpha}\widehat{\theta} - (\gamma + \frac{c+\lambda}{\alpha})^2. \tag{114}$$

Now,

$$f(\gamma_1) = 4\alpha\sqrt{\frac{F}{\alpha}}\left[2\frac{(c+\lambda)}{\alpha} - \sqrt{\frac{F}{\alpha}}\right] \geq 0, \tag{115}$$

where the last inequality follows from the condition $\frac{2(c+\lambda)}{\alpha} \geq \sqrt{\frac{F}{\alpha}}$ which defines case 3B. In addition:

$$f(\gamma_3) = 4\frac{F}{\alpha} - (\gamma + \frac{c+\lambda}{\alpha})^2 \leq 0, \tag{116}$$

where the inequality comes from the fact that square root of (86) must be positive, and

$$f(1) = 4\frac{(c+\lambda)}{\alpha}\widehat{\theta} - (1 + \frac{c+\lambda}{\alpha})^2, \tag{117}$$

which implies that, since $\hat{\theta} < 1$,

$$f(1) < 4\frac{(c+\lambda)}{\alpha} - (1 + \frac{c+\lambda}{\alpha})^2 < 0. \tag{118}$$

From (115),(116) and (118), it follows that

$$\begin{aligned} \pi_\gamma(\gamma_1) &> 0, \\ \pi_\gamma(\min\{\gamma_3, 1\}) &< 0, \end{aligned}$$

and therefore the optimal γ is attained in Region 2. The first order condition thus yields the solution.

$$4\frac{(c+\lambda)}{\alpha}\widehat{\theta} - (\gamma + \frac{c+\lambda}{\alpha})^2 = 0. \tag{119}$$

Once the appropriate root of this equation is computed, the optimal pricing schedule is obtained from Lemma 4. The last thing to derive is $q(\overline{\theta})$,

$$q(\overline{\theta}) = q(\hat{\theta}) = \frac{\gamma^2 - (\frac{c+\lambda}{\alpha})^2}{2\frac{(c+\lambda)}{\alpha}}. \tag{120}$$

and the result follows.

Proof of Proposition 4

The profit of the seller is bounded above by the total gross surplus from trade, or

$$\pi(\theta, q) \leq \alpha \left(\theta \cdot q - \frac{1}{2} q^2 \right) - (c + \lambda) q = \alpha \left[(\theta - \frac{c + \lambda}{\alpha}) q - \frac{1}{2} q^2 \right]. \tag{121}$$

The customer type which has the highest net surplus from trading is at the kink of $\widehat{s}(\theta)$, which occurs when $\theta = \sqrt{\frac{2F}{\alpha}}$.This is also the highest type for which $\widehat{s}(\theta) = 0$.

(Part: Only if) If $\sqrt{\frac{2F}{\alpha}} - \frac{c+\lambda}{\alpha} > 0$, then there is gain from trading at $\theta = \sqrt{\frac{2F}{\alpha}}$ and trade will occur.

(Part: If) If $\sqrt{\frac{2F}{\alpha}} - \frac{c+\lambda}{\alpha} < 0$,then for all $\theta \leq \sqrt{\frac{2F}{\alpha}}$, $\pi(\theta, q) < 0$, which implies it is not profitable to trade. Correspondingly, for each $\theta \in [\sqrt{\frac{2F}{\alpha}}, 1]$, the optimal surplus from on-demand computing is less than the outside opportunity of developing in-house, since the optimal surplus increases more slowly than the outside opportunity; that is,

$$\frac{d}{d\theta} \left[\frac{\alpha}{2} (\theta - \frac{c + \lambda}{\alpha})^2 \right] = \alpha (\theta - \frac{c + \lambda}{\alpha}) \tag{122}$$

and

$$\frac{d}{d\theta} \left(\frac{1}{2} \alpha \theta^2 - F \right) = \alpha \theta. \tag{123}$$

The result follows.