2011

Pricing Services Subject to Congestion: Charge Per-Use Fees or Sell Subscriptions?

Gerard. P. Cachon
University of Pennsylvania

Pnina Feldman

Recommended Citation
Cachon, G. P., & Feldman, P. (2011). Pricing Services Subject to Congestion: Charge Per-Use Fees or Sell Subscriptions?. *Manufacturing & Service Operations Management, 13* (2), 244-260. http://dx.doi.org/10.1287/msom.1100.0315

Pricing Services Subject to Congestion: Charge Per-Use Fees or Sell Subscriptions?

Abstract
Should a firm charge on a per-use basis or sell subscriptions when its service experiences congestion? Queueing-based models of pricing primarily focus on charging a fee per use for the service, in part because per-use pricing enables the firm to regulate congestion—raising the per-use price naturally reduces how frequently customers use a service. The firm has less control over usage with subscription pricing (by definition, with subscription pricing customers are not charged proportional to their actual usage), and this is a disadvantage when customers dislike congestion. However, we show that subscription pricing is more effective at earning revenue. Consequently, the firm may be better off with subscription pricing, even, surprisingly, when congestion is intuitively most problematic for the firm: e.g., as congestion becomes more disliked by consumers. We show that the absolute advantage of subscription pricing relative to per-use pricing can be substantial, whereas the potential advantage of per-use pricing is generally modest. Subscription pricing becomes relatively more attractive if consumers become more heterogeneous in their service rates (e.g., some know they are "heavy" users and others know they are "light" users) as long as capacity is fixed, the potential utilization is high, and the two segments have substantially different usage rates. Otherwise, heterogeneity in usage rates makes subscription pricing less attractive relative to per-use pricing. We conclude that subscription pricing can be effective even if congestion is relevant for the overall quality of a service.

Keywords
service operations, operations strategy, pricing and revenue management, game theory, queueing theory

Disciplines
Business Administration, Management, and Operations | Operations and Supply Chain Management | Strategic Management Policy

Pricing Services Subject to Congestion: Charge Per-Use Fees or Sell Subscriptions?

Gérard P. Cachon • Pnina Feldman

Operations and Information Management, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA

cachon@wharton.upenn.edu • pninaf@wharton.upenn.edu

June 17, 2008

Should a firm charge on a per-use basis or sell subscriptions when its service experiences congestion? Queueing-based models of pricing primarily focus on charging a fee per use of the service, in part because per-use pricing enables the firm to regulate congestion - raising the per-use price naturally reduces how frequently customers use a service. The firm has less control over usage with subscription pricing (by definition, with subscription pricing customers are not charged proportional to their actual usage), and this is a disadvantage when customers dislike congestion. However, we show that subscription pricing is more effective at earning revenue. Consequently, the firm may be better off with subscription pricing, even, surprisingly, when congestion is intuitively most problematic for the firm: e.g., as the industry moves to a standard of faster service, or as congestion becomes more disliked by consumers. We show that the absolute advantage of subscription pricing relative to per-use pricing can be substantial whereas the potential advantage of per-use pricing is generally modest. Furthermore, the relative attractiveness of subscription pricing is enhanced if the firm is able to earn third-party revenue from each transaction (e.g., if the firm acts as a platform in a two-sided market). We conclude that subscription pricing can be effective even if congestion is relevant for the overall quality of a service.

How should a firm price its service when congestion is an unavoidable reality? Customers dislike congestion, so a firm has an incentive to ensure it provides reasonably fast service. At the same time, the firm needs to earn an economic profit, so the firm's pricing scheme must generate a sufficient amount of revenue. Furthermore, these issues are closely linked: the chosen pricing scheme influences how frequently customers use a service, which dictates the level of congestion; congestion correlates with the customers' perceived value for the service, and that determines the amount of revenue the firm can generate.

A natural option is to charge customers a per-use fee or toll. Naor (1969) began this line of research and there has been many subsequent extensions of his basic model, but nearly always with a focus on per-use fees. (See Hassin and Haviv 2003 for a broad survey of this literature.)

Although the emphasis in the queueing literature has been placed on per-use pricing, other pricing schemes are observed in practice. Most notably, some firms sell subscriptions for the use of their service: a health club may charge an annual membership that allows a customer to use

the facility without additional charge for each visit; AOL, an Internet service provider, initially charged customers per-use access fees but later switched to subscription pricing (a monthly access fee with no usage limitation); Netflix, a retailer that provides movie DVDs for rental, also uses subscription pricing (a monthly fee for an unlimited number of rentals); Disney charges an entry fee for its theme park without charging per ride on the attractions; etcetera.

Despite the existence of subscriptions in practice, a subscription pricing strategy has a clear limitation in the presence of congestion effects: subscribers are not charged per use, so it is intuitive that they seek service too frequently (e.g., use the health club too often), thereby increasing congestion and decreasing the value all subscribers receive from the service. As a result, in a setting with clear congestion costs (e.g., in a queueing model) one might assume that subscription pricing would be inferior to per-use pricing. However, in this paper we demonstrate that subscription pricing may indeed be a firm's better pricing strategy despite its limitations with respect to congestion. We do so in three different capacity management scenarios: (i) the firm's service capacity is exogenously fixed; (ii) the firm's service capacity adjusts to meet an industry standard for congestion; and (iii) the firm endogenously chooses its service capacity in addition to its pricing policy.

In addition to a focus on per-use pricing, the queuing literature also assumes that revenues are earned only directly from customers, i.e., the per-use fees are assumed to be the only source of revenue. However, in some situations a firm is able to earn additional revenue from third-party sources, such as advertising revenue that is proportional to the actual use of its service (e.g., AOL). We demonstrate that the presence of such revenue favors subscription pricing relative to per-use pricing.

The next section reviews the extensive literature on pricing services, with an emphasis on models that address the issue of congestion. Section 2 details our base model. Sections 3, 4 and 5 compare the two pricing schemes under three different assumptions for how the firm's capacity is determined. Section 6 considers our model with third-party revenue. Section 7 summarizes our conclusions.

1 Related Literature

Our work is primarily related to three streams of literature: pricing in queueing models; the theory of clubs; and advance purchase pricing. Furthermore, there are some connections between our work and the relatively recent literature on two-sided markets.

Queueing theory provides a natural framework for modeling congestion, and we adopt that framework as well. However, as already mentioned, the literature on pricing of queues generally

only considers per-use pricing (e.g., Littlechild 1974, Edelson and Hilderbrand 1975, De Vany 1976, Mendelson 1985, Chen and Frank 2004). Per-use pricing is sufficient for maximizing social welfare, but it is known that a profit maximizing firm does not choose the welfare maximizing price (e.g., Naor 1969).

Randhawa and Kumar (2008) and Bitran, Rocha e Oliveira and Schilkrut (2008) do consider additional pricing schemes in queueing models. Randhawa and Kumar (2008) compare per-use pricing with a subscription pricing that imposes limits on usage, e.g., Netflix has a plan in which a customer can view as many movies as they want as long as they do not possess more than four DVDs at a time. They show that this constrained subscription plan may be better for the firm than the unconstrained per-use pricing because it reduces the volatility of the demand process the firm experiences. We do not consider subscription pricing with limitations, i.e., in our model a subscription pricing plan allows for unlimited usage. Furthermore, in their model the two plans have the same revenue potential, whereas in our model a key difference is that subscription pricing can have a higher revenue potential than per-use pricing. Hence, the restriction on usage with their subscription plan is necessary to create a distinction between the two pricing schemes. Bitran, Rocha e Oliveira and Schilkrut (2008) study a two-part tariff that combines both per-use and subscription pricing. Their focus is different than ours: they do not compare per-use to subscription pricing and instead emphasize how consumer uncertainty regarding service quality affects the dynamics of their system over time (in our model consumers have rational expectations, so we do not explicitly model the learning process).

There is a literature in economics on the pricing of shared facilities (i.e., clubs) subject to congestion, such as swimming pools and golf clubs: e.g., Berglas (1976), Scotchmer (1985). Just as in our model, customers prefer that the service/facility is used by fewer people so that there is less congestion. These papers show that a two-part tariff is optimal for the firm: a per-use fee is chosen to induce a usage level that maximizes social welfare and a subscription fee is charged to transfer all rents from customers to the firm. Like Bitran, Rocha e Oliveira and Schilkrut (2008), these papers do not compare per-use pricing to subscription pricing. Strictly speaking, according to our model the firm always prefers the two-part tariff over either subscription or per-use pricing (each is a subset of the set of two-part tariffs). However, we believe a comparison between subscription and per-use pricing is warranted. The queueing literature focuses on per-use pricing and both per-use pricing and subscriptions are observed in practice. In addition, a two-part tariff may not be desirable for reasons that we do not model (nor are generally modeled): e.g., a consumer may dislike being charged twice for the same service, especially if they do not understand the motivation

for such a pricing scheme.

Barro and Romer (1987) demonstrate that per-use pricing can be equivalent to subscription pricing. For example, they argue that a ski slope could generate the same revenue by charging a fee per ride or by charging a daily lift ticket price (which is analogous to a one-day subscription). However, in their model they assume demand exceeds the supply of ski-lift rides no matter what pricing scheme is used. Hence, a daily lift ticket price can be chosen such that usage is the same as with a per-ride price. In contrast, in our model consumers regulate their usage depending on the pricing scheme - subscription pricing leads consumers to use the facility more than any positive per-use pricing scheme. Hence, in our model the two schemes are not equivalent.

Our subscription pricing scheme resembles advance-purchase pricing (e.g., DeGraba, 1995; Xie and Shugan 2001). When consumers purchase in advance of the service, such as buying a concert ticket weeks before the event, consumers are willing to pay their expected value for the service. In contrast, when consumers spot purchase, i.e., when they know their value for the service, they are naturally willing to pay only their realized value. When purchasing in advance, consumers are more homogeneous relative to the spot market, so the firm can earn more revenue by selling in advance than by selling just with a spot price: it can be better to sell in advance to every customer at their expected value than to sell in the spot market to a portion of consumers (i.e., those consumers with a high realized value). In our model subscriptions also has this ability to extract rents because consumers are more homogeneous when they purchase subscriptions than when they purchase on a per-use basis. However, we consider the impact of congestion, whereas the advance-purchase models do not (i.e., consumers in those models do not regulate their usage based on the pricing policy).[1]

The literature on two-sided markets considers the interaction between a platform that intermediates between two markets or groups (e.g., Armstrong 2006 and Rochet and Tirole 2006). For example, a newspaper provides content to consumers and print ads to businesses. The characteristic feature of these markets is the presence of inter-market positive externalities that depend on the size of each market. For example, the value a business receives from advertising in a newspaper depends on the size of the newspaper's readership base. These models focus on the platform's pricing scheme with each market. They demonstrate that it can be optimal for the platform to subsidize one market (i.e., charge a low price, possibly zero) to generate positive externalities on the

[1] There is another difference between our model and the advance-purchasing literature. Our firm chooses a single price (either a subscription price or a per-use price). In fact, it is never optimal for the firm to offer both per-use and subscription pricing at the same time. In the advance-purchasing literature two prices are often considered (the advance price and the spot price).

other market - the lost revenue from the subsidized market is compensated by the extra revenue in the other market. This literature does not consider the negative externalities (i.e., congestion) that could be created when one market increases its size (or transaction volume), as we do. Furthermore, they do not compare subscription versus per-use pricing. We demonstrate that subscription pricing generates the benefit of a subsidy without its cost: the lack of per-use fees generates the desired positive externality on the other market while the subscription fee allows the firm to still earn revenue.

2 Model Description

A single firm provides a service to a market of potential homogenous customers of size M. Each customer finds the service to be valuable on multiple occasions, or service opportunities. For example, a customer may wish to occasionally use a teller at her bank, use the internet repeatedly or rent a movie at least a couple of times per month. This stream of service opportunities occurs for each customer at rate τ. At the moment of a service opportunity a customer observes the value, or utility, V, she would receive if she were to receive the service to satisfy that opportunity. Service values for each customer are independent and identically distributed across opportunities.[2]

Although customers value receiving the service, all else being equal, they prefer as fast a service process as possible - each customer incurs a cost w per unit of time to complete service (time waiting and in service). Hence, our model is appropriate for services that potentially exhibit varying levels of congestion. Finally, consumers neither receive utility nor incur disutility when not in the service process and waiting for the next service opportunity to arise.

The firm offers one of two pricing schemes: a per-use fee or a subscription price. The per-use fee, p, is a charge for each service completion: e.g., a fee for withdrawing money from an automatic teller machine, a fee for each visit to a health club, or a per minute fee for accessing a database. A subscription price, k, is a fee per unit of time which is independent of the amount of service the customer receives. (In our model this definition of a subscription is equivalent to a fixed fee for a finite duration with unlimited usage during that time.[3]) Where useful, we use "p" and "s" subscripts to signify notation associated with the per-use and subscription schemes, respectively.

[2] Hence, we have a single market segment of consumers, so differences between per-use and subscription pricing are not driven by a desire to price discriminate between segments, as in Essegaier, Gupta and Zhang (2002).

[3] In practice it is common to define a subscription as a fixed fee for a finite period, such as a newspaper subscription for 6 months. In our model consumers receive a steady stream of identically valued service opportunities and customers are risk neutral. As a result, any subscription defined as a fixed fee, K, for a duration, d, is equivalent in our model to a subscription rate $k = K/d$.

When a service opportunity occurs, a customer decides whether or not to seek service (i.e., join the firm's service system). The decision is based on three factors: the value of the service opportunity, the cost associated with the expected time to complete the service transaction and the firm's pricing policy. Although the customer observes the value for a particular service opportunity before deciding to seek service or not, the customer does not observe the firm's current queue length. However, the customer has an expectation for the average arrival rate of customers to the firm's service, λ, and the customer knows the function that translates an arrival rate into an expected service time, $W(\lambda)$.[4] We use the term *service time* to refer to the total time to complete the service, i.e., it includes time waiting and in service. The function $W(\lambda)$ exhibits the following natural properties: $W'(\lambda) > 0$ and $W''(\lambda) \geq 0$. Thus, $wW(\lambda)$ is the expected cost to the customer of the time to receive one service opportunity. We refer to $wW(\lambda)$ as the service time cost or the congestion cost. Note, a customer cannot balk (or, chooses not to balk) from the queue after choosing to seek service (otherwise, the customer would effectively be able to observe the queue length before the joining decision is made).[5] Finally, the firm's pricing policy clearly influences the customer's decision. With each service opportunity the customer decides whether to seek service based on the amount of utility that would be earned from the opportunity relative to congestion costs and the firm's per-use fee (which in the case of subscription pricing, is zero). Whether to adopt a subscription is based on the expected arrival of service opportunities and their expected net utilities. Consumers are risk neutral and make choices based on the average utility each option generates (rather than the discounted utility of each option). In addition, consumers make pure strategy choices (join the service system or not, subscribe or not). Allowing mixed strategy choices either favors subscription pricing or has no impact on our results.[6]

[4] This is actually a stronger informational assumption than is needed. We merely require the customer to have an expectation of the firm's service time and that expectation must be correct (i.e., they do not need to know the relationship between the service time and λ).

[5] We suspect our qualitative results continue to hold even if balking is allowed. In that case, subscription consumers join only if the queue length is sufficiently small so that the value from the service exceeds the expected time costs. Per-use customers make essentially the same decision, but they compare their net utility (value minus per-use fee) to the expected waiting cost. Hence, for any queue length, the arrival rate to the system with subscription pricing will be no less than with per-use and generally will be strictly greater. Therefore, subscription pricing still leads to more congestion than per-use pricing. Furthermore, consumers are still homogeneous when deciding whether to purchase a subscription or not (because they base their decision on expected usage and system time costs). We do not work with the balking model because it is analytically more cumbersome - the system time function, $W(\cdot)$ is analytically complex and probably depends on the particular pricing scheme in use (whereas in our model it depends only on the arrival rate, λ).

[6] Consumers need to decide with each service opportunity whether to seek service or not. The optimal strategy for a consumer is always a pure strategy conditional on the value of the service opportunity. Hence, including mixed strategies has no impact with this decision. Regarding the subscription decision, we find that the firm's profit can be higher if mixed strategies are allowed in the exogenous capacity model. However, the firm's profit is unchanged in the other two capacity models by the inclusion of mixed strategies.

The server's processing rate is μ. (Thus, $W(0) = 1/\mu$ because $1/\mu$ is a customer's service time when there is no congestion.) In section 3 we assume μ is exogenous, whereas in section 4 the firm adjusts μ to meet an exogenously set standard for service time and in section 5 the firm chooses μ subject to a fee that is proportional to the service rate. Naturally, $W(\lambda)$ is decreasing in μ. Furthermore, we assume $W(\tau M)$ is sufficiently small relative to $1/\tau$, where $\Lambda = \tau M$ is the maximum possible arrival rate of service opportunities (i.e., the arrival rate when every customer seeks service at every service opportunity). This implies that for a fixed potential arrival rate of service, Λ, the potential population of customers, M, is large and they do not seek service too frequently (τ is small). Consequently, the arrival rate to the firm's queue is (approximately) independent of the queue length (which is typically assumed in the queueing literature) and there is little chance that a service opportunity arises while a customer is in the service process. For example, a customer does not receive another need to withdraw cash from an automatic teller machine while the customer is in the process of withdrawing cash.[7]

To complete the definition of the model, we provide some additional structure for the service value distribution and the system-time function. Let $F(\cdot)$ be the distribution function and $f(\cdot)$ the density function of each service value: assume F is differentiable, $F(0) = 0$, and F exhibits an increasing failure rate (IFR). For some results we invoke one of the following additional assumptions related to the hazard rate, $h(x) = f(x)/\bar{F}(x)$, where $\bar{F}(x) = 1 - F(x)$:

Assumption 1 *(A1)* $h'(x)/h(x)^2$ *is decreasing*

Assumption 2 *(A2)* $xh'(x)$ *is increasing*

(A2) holds for a power distribution with parameter $\kappa > 1$, while both (A1) and (A2) hold if F is uniform on the support $[0, \overline{v}]$ or Weibull with parameters $\kappa \geq 1$ and $\beta > 0$. (Note, a Weibull distribution with $\kappa = 1$ is an exponential distribution.) In all three versions of the model, we assume F is uniform on the support $[0, \overline{v}]$ to derive analytical comparisons between the pricing schemes. Regarding the system-time function, for the industry standard model (section 4) and in the capacity choice model (section 5), we assume $W(\lambda) = 1/(\mu - \lambda)$, which corresponds to the expected time in an $M/M/1$ queue with first-come-first serve priority. Furthermore, we use that functional form to compare the pricing schemes in the exogenous capacity model (section 3). The electronic companion provides details for results we claim in this text without explicit proof

[7] See Randhawa and Kumar (2008) for a model of a closed queueing system in which a consumer's service opportunity process is "turned off" when the consumer is in the service system. Consequently, in their model the arrival rate to the queue depends on the number of customers in queue.

or support. Some of our analytical results can be obtained under less restrictive distributional assumptions, and these are noted in the electronic companion. Furthermore, the electronic companion provides numerical evidence that our results generalize beyond the specific assumptions we adopt for analytical tractability.

3 Exogenous Capacity

In this section we analyze a version of our model in which the firm's service processing rate, μ, or capacity, is exogenously fixed with either pricing scheme. This analysis is appropriate for a firm that has the short term flexibility to modify its pricing but does not have the short term ability to alter its capacity. For each pricing scheme we derive the firm's equilibrium arrival rate and optimal revenues, which allows us to establish conditions under which one scheme is preferred over another.

3.1 Per-Use Pricing

With per-use pricing a customer observes the realized value of a particular service opportunity and then requests service if the net utility is non-negative, i.e., the value of that opportunity is greater than or equal to $p + wW(\lambda)$.[8] Given that p, w and λ are common to all customers (they all have the same expectations) and constant across time, there is some threshold value, v, such that a customer seeks service whenever the realized value of an opportunity is v or greater, and otherwise the customer passes on the opportunity:

$$v = p + wW(\lambda).$$

The actual arrival rate to the service is then $\Lambda \bar{F}(v)$. For expectations to be consistent with actual operating conditions (i.e., $\lambda = \Lambda \bar{F}(v)$) the threshold v must satisfy

$$v = p + wW\left(\Lambda \bar{F}(v)\right). \tag{1}$$

Given that W is increasing, it follows that there is a unique solution to (1). Furthermore, the threshold is increasing in the per-use fee, p.

[8] In some situations it is reasonable to suspect that the waiting cost function is not constant, but it depends on the value that the consumer attaches to the service opportunity. A consumer may find waiting more or less costly as she values the service more. Most of the results in this section generalize for a linear waiting cost function (i.e. $w(v) = a + bv$, where b can be either positive or negative).

The firm's revenue is $R_p = \lambda p$, which can be expressed in terms of the threshold v :

$$R_p(v) = \Lambda \bar{F}(v) \left(v - wW(\Lambda \bar{F}(v)) \right).$$

The following theorem establishes that an optimal threshold, v_p, exists and is unique (with this and the subsequent theorems, see the appendix for the proofs).

Theorem 1 *The per-use revenue function, $R_p(v)$, is quasi-concave and $v_p = \arg\max_v R_p(v)$ is uniquely defined by*

$$v_p = \frac{\bar{F}(v_p)}{f(v_p)} + wW(\Lambda \bar{F}(v_p)) + w\Lambda \bar{F}(v_p) W'(\Lambda \bar{F}(v_p)). \tag{2}$$

To translate v_p back into an actual price, the firm's optimal per-use fee is

$$p_p = \frac{\bar{F}(v_p)}{f(v_p)} + w\Lambda \bar{F}(v_p) W'(\Lambda \bar{F}(v_p)).$$

3.2 Subscription Pricing

With a subscription scheme there is no explicit fee charged per transaction, e.g., the members of a health club can use the service whenever they wish without additional charge. However, a customer may not take advantage of a service opportunity if her value for that opportunity is low relative to her expectation of congestion costs, and that expectation depends on the number of subscribers and the frequency of their usage. For now, we assume all consumers subscribe and then we confirm that expectation is correct. As a result, if each consumer uses the threshold v_s to decide whether to seek service or not, then the arrival rate to the service is $\Lambda \bar{F}(v_s)$: Λ is the arrival rate of service opportunities conditional that all M consumers are subscribers and $\bar{F}(v_s)$ is the fraction of service opportunities that generate a service request. In equilibrium, the indifferent consumer's value, v_s, exactly equals the expected congestion cost:

$$v_s = wW\left(\Lambda \bar{F}(v_s)\right). \tag{3}$$

Now consider whether to purchase a subscription or not. At the time this decision is made the customer does not know when future service opportunities will occur or their values, but does know his/her threshold value, v_s, for seeking service. Hence, as part of the purchasing decision, a customer expects that a subscription generates the following net value per service opportunity,

$$\bar{F}(v_s) \left(E\left[V | V \geq v_s\right] - v_s \right) :$$

$\bar{F}(v_s)$ is the probability a service opportunity is sufficiently valuable to seek service, $E[V|V \geq v_s]$ is the value received conditional that a service opportunity yields a value greater than the threshold and the last term, v_s, is the expected congestion cost (from (3)).

Given that service opportunities arrive at rate τ, it is optimal for the firm to set the subscription rate, k, equal to the value of a subscription per unit of time (net of system-time cost)[9]:

$$k = \tau \bar{F}(v_s) \left(E[V|V \geq v_s] - v_s \right).$$

All consumers purchase a subscription even though they are indifferent between doing so or not, which confirms our initial assumption that all consumers subscribe.[10] As a result, subscription pricing allows the firm to extract all consumer surplus, conditional on the level of congestion that subscriptions generate.[11]

The firm's resulting revenue can be expressed in terms of the threshold v_s :

$$R_s(v_s) = kM = \Lambda \bar{F}(v_s) \left(E[V|V \geq v_s] - v_s \right).$$

Note, while the threshold v_p was a decision variable for the firm with per-use pricing, the firm has no control over the threshold v_s with subscription pricing - it is set by (3). In other words, with subscription pricing (and exogenous capacity) the firm cannot control congestion, even though it possesses an effective mechanism for maximizing revenue conditional on the system's congestion.

3.3 Comparison between Pricing Schemes

This section compares the revenues generated by the two pricing schemes. To make these comparisons more explicit, we assume in this section $V \sim U[0, \overline{v}]$ and $W(\lambda) = 1/(\mu - \lambda)$.

Per-use pricing allows the firm to control congestion by regulating the service arrival rate, but the per-use fee must also earn rents for the firm. In contrast, subscription pricing is weak with respect to controlling congestion, but does allow the firm to extract rents efficiently. The firm's preference between these two schemes, therefore, depends on the relative strength of these two countervailing factors.

[9] Lowering k merely reduces revenue per customer without changing demand, so that cannot be optimal. There is no demand with a higher k, so that is not optimal either.

[10] If consumers can adopt a mixed strategy with respect to the subscription purchase decision, then the firm may be able to earn a higher profit by charging an even higher subscription rate. If it does so, then each consumer purchases a subscription with some probability, say γ, so that the expected arrival rate is $\gamma\Lambda\bar{F}(v_s)$ and $v_s = wW\left(\gamma\Lambda\bar{F}(v_s)\right)$. Hence, our results provide a lower bound on the profit with subscription pricing.

[11] As discussed in section 1, in advance selling models the firm extracts all consumer surplus with the advance price. However, in those models the potential consumer surplus is independent of the pricing scheme, whereas here the amount is not (it depends on how much congestion materializes).

We now define the set of parameters for which the firm can earn non-negative revenue. Although the firm's problem is determined by four parameters (w, μ, Λ and $\overline{v}$), the next theorem indicates that the pricing schemes' relative rankings depend only on two of them.

Lemma 1 *The optimal revenue with each pricing scheme (R_p and R_s) can be expressed in terms of α, ρ, Λ and $\overline{v}$, where $\alpha = w/\mu\overline{v}$ and $\rho = \Lambda/\mu$. Furthermore, the two revenues are non-negative for $\alpha \in [0,1]$ and linear in $\Lambda\overline{v}$. Consequently, the relative revenue, R_s/R_p, can be expressed in terms of α and ρ.*

The term $\Lambda\overline{v}$ merely scales the revenues, so it does not influence their relative rankings. Instead, whether per-use pricing or subscriptions are preferred depends on α (which measures the relative strength of congestion costs to service values) and the potential utilization rate of the system, ρ.

Theorem 2 *For each value of α, there exists a unique $\widetilde{\rho}\,(\alpha)$, such that subscription yields higher revenue than per-use pricing for $\rho < \widetilde{\rho}(\alpha)$ (recall, ρ is the potential utilization, Λ/μ). Otherwise, per-use pricing yields higher revenue. Moreover, $\widetilde{\rho}(\alpha)$ is decreasing in α.*

From Theorem 2, per-use is preferred over subscription for highly congested systems. The key issue is the degree of congestion needed for per-use to be preferred. For various levels of congestion costs, α, Table 1 provides the potential utilization rate, $\widetilde{\rho}(\alpha)$, at which the two schemes yield the same revenue. It can be demonstrated that $\lim_{\alpha\to 0}\widetilde{\rho}(\alpha) = \sqrt{2}$ and $\lim_{\alpha\to 1}\widetilde{\rho}(\alpha) = 1$. Thus, subscription pricing always generates higher revenue than per-use pricing when the potential arrival rate to the queue is less than the processing rate. Subscription pricing can be preferred even if the potential arrival rate is as much as 140% of the firm's processing rate. Subscription pricing can also be preferred when the system's actual utilization rate is high. Table 1 lists the system's actual utilization rate when the potential utilization rate is $\widetilde{\rho}(\alpha)$. For example, when $\alpha = 0.01$ and $\Lambda = 1.411\mu$, subscription pricing yields the same revenue as per-use pricing even though the actual utilizations are 96.8% and 64.8% respectively. In addition, the actual utilization rates are increasing in ρ. Thus, when $\alpha = 0.01$, subscription pricing is preferred whenever it yields an actual utilization rate that is lower than 96.8%. Hence, although subscription pricing cannot control congestion well, it still generates higher revenue than per-use pricing even in systems with a considerable amount of congestion.

To explore the strength of subscription pricing further, the next theorem characterizes revenues with extreme levels of potential congestion.

Table 1. Potential utilization rates, $\widetilde{\rho}(\alpha)$, that yield identical revenue with per-use and subscription pricing, as well as actual utilizations when the potential arrival rate is $\widetilde{\rho}(\alpha)\mu$.

		Actual utilization (%) when $\rho = \widetilde{\rho}(\alpha)$	
α	$\widetilde{\rho}(\alpha)$	Per-use	Subscription
0.99	1.002	0.3	0.5
0.75	1.069	7.1	13.8
0.50	1.153	16.4	31.3
0.25	1.264	30.5	55.5
0.01	1.411	64.8	96.8

Theorem 3 *The following limits hold: (i)* $\lim_{\rho\to 0} R_s = \Lambda\overline{v}(1+\alpha)^2/2$ *and* $\lim_{\rho\to 0} R_p = \Lambda\overline{v}(1+\alpha)^2/4$. *(ii)* $\lim_{\rho\to\infty} R_s/R_p = 0$ *and* $\lim_{\rho\to\infty} R_x = 0$, $x \in \{s,p\}$.

Subscription pricing generates twice as much revenue as per-use when capacity is unlimited ($\rho = 0$). Therefore, subscription pricing starts with a considerable advantage relative to per-use pricing. As a result, congestion needs to be substantial in the system before the congestion-controlling benefits of per-use pricing dominates the rent-extracting capability of subscription pricing. Furthermore, revenue declines in ρ with all schemes, so per-use pricing dominates subscription pricing only when revenues are in fact low. This suggests that per-use pricing can provide only a modest absolute advantage relative to subscription pricing, but the absolute advantage of subscription pricing can be substantial. Taken together, these results indicate that from a practical perspective, subscription pricing can indeed be better than per-use pricing even if capacity is fixed and the system is subject to congestion related costs.

4 Industry Standard for Service Time

In this section we consider a model in which the firm must conform to a predetermined industry standard for service time. For example, in the call-center industry it is common to set a standard in terms of the probability that a customer's wait will not exceed a certain amount of time (Gans *et al.* 2003; Cleveland and Mayben 1997). We want to determine how the presence of an industry standard influences the relative performance of the pricing schemes we study.

We continue to work with the assumption that the firm's service process can be well approximated by an $M/M/1$ queue, i.e., $W(\lambda) = 1/(\mu - \lambda)$. In that case, let T be the time a customer

spends in the service process (again, waiting and in service), and let t be a benchmark time:

$$\Pr\{T \leq t\} = 1 - e^{-(\mu-\lambda)t} \tag{4}$$

Define the industry standard to be that customers are in the system no more than t units of time with at least probability θ. From (4), the firm must have the following minimum capacity to achieve this standard:

$$\mu = \lambda + \frac{\log(1-\theta)}{t},$$

i.e., the firm must have enough capacity to process its arrival rate plus a fixed buffer that only depends on the industry standard. For notational convenience, define $I = \mu - \lambda$, so I is the size of that buffer and an increase in I implies an increase in the standard.

We assume the firm adjusts its capacity to meet the standard, but each unit of capacity costs the firm c per unit of time. Therefore, to the extent that the firm's arrival rate of customers varies between pricing schemes, so does its capacity, as already mentioned.

4.1 Per-Use Pricing

As in the fixed capacity model, with per-use pricing a customer with value v is indifferent between requesting service or not, where $v = p + w/\left(\mu - \Lambda\overline{F}(v)\right) = p + w/I$. The firm's profit function, assuming the arrival rate is strictly positive, in terms of this threshold, is

$$\begin{aligned}\Pi_p(v) &= \Lambda\bar{F}(v)p - c\mu(v) \\ &= \Lambda\bar{F}(v)\left(v - \frac{w}{I} - c\right) - cI,\end{aligned}$$

where $\mu(v) = \Lambda\overline{F}(v) + I$ is the firm's capacity. The following theorem establishes that an optimal threshold, v_p, exists and is unique.

Theorem 4 *The per-use profit function $\Pi_p(v)$ is quasi-concave and $v_p = \arg\max_v \Pi_p(v)$ is uniquely defined by*

$$v_p = \frac{\bar{F}(v_p)}{f(v_p)} + \frac{w}{I} + c \tag{5}$$

As the standard increases, more customers use the service (v_p decreases) even though the per-use price, $p(v_p) = \bar{F}(v_p)/f(v_p) + c$, increases.

Due to the fixed cost, cI, the firm may not earn a positive profit, i.e., $\Pi_p(v_p) < 0$ is possible. If (A1) holds, we find that $\Pi_p(v_p)$ is convex-concave in I. In such case, profit is negative if the standard is too low or too high. If the standard is too low, the service is of poor quality so few customers

use it sparingly, thereby generating too little revenue. If the standard is too high, customers use the service extensively, but at a decreasing rate, so the incremental revenues cannot cover the large cost of the necessary buffer capacity. Positive profit requires an intermediate standard to provide enough demand without an excessive capacity cost.

4.2 Subscription Pricing

There exists a threshold, v_s, as in the fixed capacity model, such that all consumers with value v_s or higher seek service, $v_s = w/\left(\mu_s - \Lambda\overline{F}\left(v_s\right)\right)$, conditional that all consumers are subscribers. To conform to the industry standard, capacity adjusts so that $\mu(v_s) = \Lambda\overline{F}\left(v_s\right) + I$, which implies $v_s = w/I$ in equilibrium. The subscription rate is set so that all consumers purchase a subscription (which confirms the initial assumption that all are subscribers) and the firm's profit is then

$$\begin{aligned}\Pi_s(v_s) &= \Lambda\overline{F}\left(v_s\right)\left(E\left[V|V \geq v_s\right] - v_s\right) - c\mu(v_s) \\ &= \Lambda\overline{F}\left(v_s\right)\left(E\left[V|V \geq v_s\right] - v_s - c\right) - cI\end{aligned}$$

There is no guarantee that $\Pi_s(v_s) \geq 0$. As with per-use pricing, if (A1) holds, then $\Pi_s(v_s)$ is convex-concave in I - profits are positive only with intermediate levels of the standard.

4.3 Comparison

As in the fixed capacity model, we assume $V \sim U\left[0, \overline{v}\right]$ in this section to compare subscription and per-use pricing. Subscription pricing results in a higher arrival rate than per-use pricing (i.e., a lower threshold, $v_s = w/I < v_p$) and requires more capacity ($\mu(v_s) > \mu(v_p)$) to ensure the standard is met. The next theorem establishes a useful comparative result.

Theorem 5 *If* $I > w/\left(\overline{v} - c\right)$, *then* $\partial\Pi_p/\partial I < \partial\Pi_s/\partial I$.

The condition in Theorem 5, $w/I < \overline{v} - c$, is necessary for profit to be positive with either scheme, but not sufficient: with either pricing scheme profit is negative when $I \leq w/\left(\overline{v} - c\right)$. Therefore, the industry standard must be sufficiently high for either scheme to earn a positive profit (but not too high). Furthermore, increasing the standard favors subscription pricing relative to per-use pricing in the sense that $\Pi_s - \Pi_p$ is increasing in I.[12] In fact, $\Pi_s > \Pi_p$ whenever

$$\frac{w}{\overline{v} - \left(\sqrt{2} + 1\right)c} \leq I : \tag{6}$$

[12] Theorem 5 generalizes for all distributions that satisfy (A1) by requiring $\overline{F}\left(v_p\right) < \overline{F}\left(v_s\right)\left(1 - ch\left(v_s\right)\right)$ instead of $I > w/\left(\overline{v} - c\right)$ (the latter condition is equivalent to the former for the uniform distribution). It can be shown that for all distributions that satisfy (A1), there exists a unique $\widetilde{I}$ such that $\partial\Pi_p/\partial\widetilde{I} = \partial\Pi_s/\partial\widetilde{I}$. Hence, a higher industry standard favors subscription under broader conditions.

less expensive capacity and lower sensitivity to congestion favor subscription pricing, which is intuitive. However, (6) also indicates that a higher industry standard favors subscription, which is surprising: it might seem that a pricing structure that is better at controlling congestion would be preferable when service time standards are more strict. To explain, note that when the standard is very high, most of the capacity is buffer capacity, i.e., $\mu \approx I$ (or, put another way, $I >> \lambda$). Hence, with a high standard, the two schemes incur nearly the same capacity cost. Moreover, with a high standard, service times are nearly inconsequential because there is essentially no congestion. Without congestion, the revenue rate equals $\bar{v}\Lambda/2$ under subscription pricing and $\bar{v}\Lambda/4$ under per-use pricing. Therefore, subscription is more likely to be profitable when the industry standard is high – its revenue rate is much higher so it is more likely to cover the large fixed cost of excess capacity.[13]

5 Capacity Choice

In section 3 the firm can choose how to price but not its capacity, so the pricing decision results only in variation in service time. In section 4 the firm can choose how to price but not the service time it delivers, so the pricing decision results only in variation in capacity. In this section the firm chooses how to price and its capacity, so the pricing decision influences both the firm's capacity and its service time. As in section 4, capacity is expensive - the firm incurs a cost at rate $c\mu$ for maintaining capacity μ, where $c > 0$. Furthermore, we continue to assume $W(\lambda) = 1/(\mu - \lambda)$.

5.1 Per-Use Pricing

The consumer's choice in this setting is the same as in the fixed capacity model. As a result, we can express the firm's profit function in terms of the threshold of the indifferent consumer, v, and capacity:

$$\begin{aligned} \Pi_p(v,\mu) &= R_p(v) - c\mu \\ &= \Lambda \bar{F}(v)\left(v - \frac{w}{\mu - \Lambda\bar{F}(v)}\right) - c\mu. \end{aligned}$$

[13] The comparison between Π_s and Π_p is more complex if one restricts attention to positive profit. It is possible that subscription pricing is profitable for some standards whereas per-use pricing is not profitable for all standards. The opposite is possible as well: $\Pi_s < 0$ for all standards while $\Pi_p > 0$ for some standards. Finally, it is also possible that $\Pi_s < 0 < \Pi_p$ for $I < I'$, $0 < \Pi_s < \Pi_p$ for $I' < I < I''$, $0 < \Pi_p < \Pi_s$ for $I'' < I < I'''$, and $\Pi_p < 0 < \Pi_s$ for $I''' < I < I''''$. However, even in this analysis, subscription pricing tends to be more profitable as the standard increases.

The profit function is concave in μ, so it is straightforward to determine that $\mu_p(v)$ is the firm's optimal capacity for a given threshold, v, where,

$$\mu_p(v) = \Lambda \bar{F}(v) + \sqrt{\frac{w\Lambda \bar{F}(v)}{c}}.$$

(Note, although our notation is similar, this capacity function is different than in the industry standard model, section 4.) The firm's profit rate is then

$$\begin{aligned} \Pi_p(v) &= \Pi_p\left(v, \mu_p(v)\right) \\ &= \Lambda\left(\bar{F}(v)(v-c) - 2\phi\sqrt{\bar{F}(v)}\right) \end{aligned}$$

where the constant ϕ is defined for convenience:

$$\phi = \sqrt{cw/\Lambda}.$$

The following theorem establishes the uniqueness of the optimal per-use threshold.

Theorem 6 *If $\bar{v} > c$, there exists an upper bound $\overline{\phi}_p$, such that for every $\phi \leq \overline{\phi}_p$ there exists a unique optimal threshold, $v_p = \arg\max_v \Pi_p(v)$ that yields positive profit, $\Pi_p(v_p) \geq 0$. This threshold is the smallest solution to the implicit equation given by:*

$$v_p = \frac{\bar{F}(v_p)}{f(v_p)} + \frac{\phi}{\sqrt{\bar{F}(v_p)}} + c. \tag{7}$$

Furthermore, if (A1) holds, then there exist two solutions to (7) and the smallest solution is the unique optimal threshold. The optimal capacity is

$$\mu_p = \Lambda \bar{F}(v_p) + \frac{\phi\Lambda\sqrt{\bar{F}(v_p)}}{c} \tag{8}$$

and the firm's per-use fee is

$$p_p = v_p - w/\left(\mu_p - \Lambda\bar{F}(v_p)\right).$$

The bound in Theorem 6, $\overline{\phi}_p$, merely states that the firm can earn a positive profit only if capacity is sufficiently cheap, customers are sufficiently patient and the market is sufficiently large.

5.2 Subscription Pricing

With subscription pricing and a fixed capacity the firm has little control over congestion. However, the firm gains some control over congestion when the firm can choose its capacity. In particular,

if μ is the firm's capacity, then a consumer with value $v = w/\left(\mu - \Lambda\bar{F}(v)\right)$ is indifferent between seeking service or not. Instead of thinking in terms of the firm choosing μ, we can use that relationship to frame the firm's problem in terms of choosing the threshold, v,

$$\mu_s(v) = \Lambda\bar{F}(v) + w/v$$

The firm's profit function can then be written as

$$\Pi_s(v) = \Lambda\bar{F}(v)\left(E[V|V \geq v] - v\right) - c\left(\Lambda\bar{F}(v) + w/v\right)$$

where the first term is the revenue the firm earns from subscriptions assuming the firm chooses the maximum subscription fee that induces all consumers to purchase a subscription, conditional on the expected level of congestion.

Theorem 7 *If $E[V] > c$, there exists an upper bound $\overline{\phi}_s$, such that for every $\phi \leq \overline{\phi}_s$, there exists an optimal threshold, $v_s \in \arg\max \Pi_s(v)$, that yields positive profit, $\Pi_s(v_s) \geq 0$. That threshold is implicitly defined by:*

$$v_s = \phi\sqrt{\frac{1}{\bar{F}(v_s) - cf(v_s)}}. \tag{9}$$

Furthermore, if (A2) holds, then there exist two solutions to (9) and the smallest solution is the unique optimal threshold.

As with Theorem 6, Theorem 7 indicates that a positive profit occurs only when capacity is not too expensive, customers do not incur time costs that are too high and there is a sufficient number of customers in the market. However, the two bounds, $\overline{\phi}_p$ and $\overline{\phi}_s$ need not be the same.

5.3 Comparison

In this section we assume $V \sim U[0, \overline{v}]$, but we observe numerically that these results hold for the Weibull distribution with $\kappa \geq 1$.

As with a fixed capacity, it is possible to show that per-use pricing leads to a system with less congestion than subscription pricing: $v_s < v_p$. The firm invests more in capacity with subscription pricing (to control congestion somewhat) than with per-use pricing: $\mu_p < \mu_s$. Even though the firm invests more in capacity with subscription pricing, congestion is also higher with that scheme: $u_s(c) > u_p(c)$, where $u_x(c)$ is the actual utilization rate,

$$u_x(c) = \Lambda\bar{F}(v_x)/\mu_x, \qquad x \in \{s, p\}.$$

If capacity is inexpensive, $c = 0$, subscription pricing performs strictly better than per-use pricing,

$$\Pi_s(v_s|c=0) > \Pi_p(v_p|c=0) :$$

without the concern of congestion, the revenue extraction benefit of subscription pricing dominates. However, subscription profits decrease at a faster rate with respect to the cost of capacity,

$$\frac{\partial \Pi_s(v_s)}{\partial c} < \frac{\partial \Pi_p(v_p)}{\partial c} < 0 :$$

subscription pricing is more sensitive to capacity costs than per-use pricing. Define $\bar{c}_x$, as the maximum capacity cost that allows a non-negative profit with pricing scheme $x \in \{s, p\}$. Combining these results, one of two scenarios emerges: either $\bar{c}_s \leq \bar{c}_p$ or $\bar{c}_p \leq \bar{c}_s$.

Consider the first scenario, $\bar{c}_s \leq \bar{c}_p$. There exists some $\tilde{c}$ such that the two schemes earn the same profit, $\Pi_s(v_s|\tilde{c}) = \Pi_p(v_p|\tilde{c}) > 0$. It follows that subscription yields higher profit than per-use pricing for $c \in [0, \tilde{c}]$ while per-use is better for $c \in [\tilde{c}, \bar{c}_p]$. Furthermore, for $c \in [\bar{c}_s, \bar{c}_p]$, subscription pricing cannot earn a positive profit whereas per-use pricing does. That is what one might expect given that subscription pricing gives the firm less control over congestion - if capacity costs are sufficiently high, per-use pricing is preferable and may be the only scheme that yields a positive profit.

Now consider the second scenario, $\bar{c}_p \leq \bar{c}_s$. Subscription pricing is preferred if $c \in [0, \bar{c}_p]$ and subscription pricing is the only scheme that returns a positive profit if $c \in [\bar{c}_p, \bar{c}_s]$. In other words, it is possible that subscription pricing is the preferred scheme for any capacity cost that allows the firm to make a profit. Furthermore, if capacity is sufficiently expensive, it is possible that subscription pricing can yield a profit whereas per-use pricing cannot: in those situations capacity is sufficiently expensive that per-use pricing is unable to extract enough revenue from customers to cover the cost of capacity.[14]

Figure 1 illustrates these results. The left hand graph corresponds with the first scenario and the right hand graph corresponds to the second scenario. We note that subscription pricing performs better than per-use if the capacity cost is low. Furthermore, while per-use pricing can be more profitable than subscription pricing, it is only more profitable when capacity is sufficiently expensive. As a result, the absolute advantage of per-use pricing is generally small, whereas the absolute advantage of subscription pricing can be large.

[14] This result provides an interesting contrast with the necessary conditions for each pricing scheme to be profitable. Recall, $E[V] > c$ is necessary for subscription pricing while the less restrictive $\overline{v} > c$ is necessary for per-use pricing. These are only necessary conditions, and not sufficient conditions, as we have demonstrated. Therefore, it would be

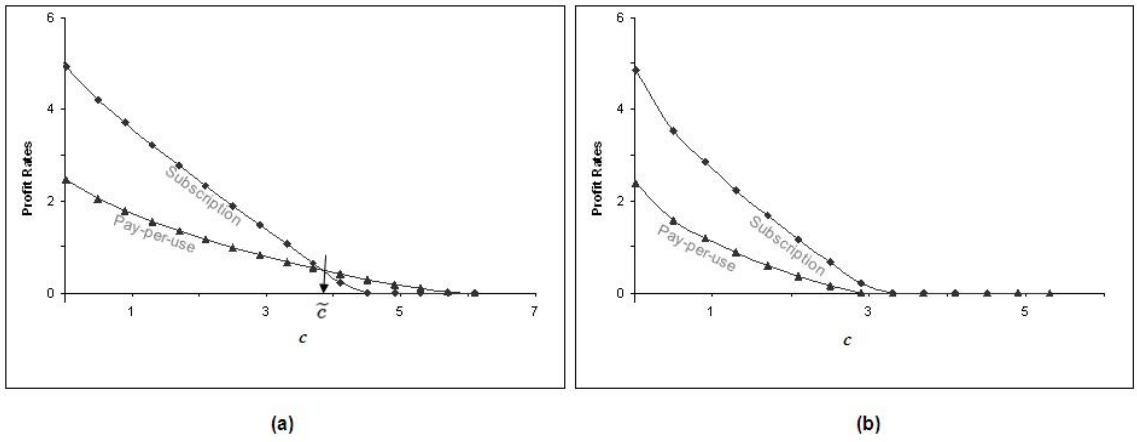

Figure 1. Profit rates of the two pricing schemes with respect to the capacity cost, c. The following parameter values are used: (a) $w = 0.05$; and (b) $w = 0.5$. ($\Lambda = 1$ and $\overline{v} = 10$ in both panels.)

The only difference between the two panels in Figure 1 is that the right hand side has a higher waiting cost: $w = 0.5$ instead of $w = 0.05$. In fact, it can be shown that there exists a $\widetilde{w}$ such that for all $w > \widetilde{w}$ the second scenario occurs, i.e., if the waiting cost is sufficiently high, subscription pricing dominates per-use pricing for all capacity costs that yield a positive profit. In other words, when congestion is most costly, in the sense that the service-time cost is high, then subscription pricing can be better than per-use pricing even though it has less control over congestion. This counter-intuitive result is similar to our finding for the industry standard model - if congestion costs are high, a large capacity must be chosen to minimize congestion, and this can only be profitable when the pricing scheme is able to extract a sufficient amount of revenue.

It is also illustrative to compare the pricing schemes with respect to utilization. It can be shown that the relevant metric is $w/\Lambda\overline{v}$. (Note, with a fixed capacity we use $w/\mu\overline{v}$ for making comparisons between the two pricing schemes, but now μ is endogenous and different across schemes.) Table 2 provides the firm's utilization under each pricing scheme when capacity is $\widetilde{c}$, i.e., when the cost of capacity is such that per-use and subscription pricing yield the same profit. (If w were any higher, then subscription pricing dominates per-use pricing for all utilizations that yield a positive profit, i.e., in that case we enter the $w > \widetilde{w}$ regime.) We observe numerically that utilization is increasing in c with each pricing scheme. Consequently, subscription pricing is better than per use pricing for all utilizations that are lower than those indicated in the table. For example, when $w/\Lambda\overline{v} = 0.03$, subscription pricing is better than per-use pricing whenever it yields a utilization of 80% or lower. The table indicates that subscription pricing can be better than per-use pricing

misleading to conclude from those conditions that a high capacity cost favors per-use pricing in all circumstances.

Table 2. Utilization when capacity is such that subscription pricing and per-use pricing yield the same profit (i.e., $c = \tilde{c}$).

$w/\Lambda\overline{v}$	u_p	u_s
0.030	61.35%	80.19%
0.025	64.05%	81.87%
0.020	67.16%	83.67%
0.010	75.43%	88.36%
0.0050	81.87%	91.72%
0.0025	86.77%	94.12%
0.0020	88.07%	94.75%
0.0010	91.38%	96.26%
0.0005	93.82%	97.35%
0.0002	96.04%	98.32%
0.0001	97.18%	98.81%
0.00005	97.99%	99.16%
0.00001	99.10%	99.62%

even if the utilization rate is quite high (say, higher than 98%). Therefore, as in the previous two models, subscription pricing can be better than per-use pricing even if it results in a highly utilized system.

6 Extensions: Third-Party Revenue

Just as in the queuing literature with pricing, our base model assumes revenue is earned from customers and only directly from customers. However, there are services that are able to generate revenue indirectly from their customers. For example, AOL can charge for its internet service but also can collect revenue from firms based on its customer's transactions. Other firms, such as Yahoo and Google, do not charge customers for their on-line search services, but do collect revenue from firms based on those searches (e.g., Google's AdWords revenue, Auchard 2007). The purpose of this section is to explore how the relative merits of per-use and subscription pricing compare in the presence of third-party revenue (i.e., revenue that is not earned from customers but rather from other sources based on customer usage of the service).

Consider our fixed capacity model (section 3) with $W(\lambda) = 1/(\mu - \lambda)$, but now the firm earns r per transaction from sources other than customers.[15] The inclusion of r has no direct impact on customers, so their decisions remain unchanged with respect to whether to purchase a subscription

[15] Alternatively, a firm may generate third-party revenue that is proportional to the number of customers. For example, retail banks collect deposits from their customers and those deposits generate revenue. In our model all customers use the service with either pricing scheme, so this source of revenue does not have an impact on the analysis of our model.

and on which service opportunities to use the service. Thus, the per-use revenue function is

$$R_p(v) = \Lambda \bar{F}(v) \left(v - \frac{w}{\mu - \Lambda \bar{F}(v)} + r \right),$$

where v is the customers' threshold for seeking service or not. The firm's optimal threshold, v_p, is now implicitly defined by

$$v_p + r = \frac{\bar{F}(v_p)}{f(v_p)} + \frac{w\mu}{\left(\mu - \Lambda \bar{F}(v_p)\right)^2}.$$

The subscription revenue function is:

$$R_s(v) = \Lambda \bar{F}(v) \left(E\left[V | V \geq v\right] - v + r \right).$$

The optimal threshold, v_s, as before, is given by

$$v_s = \frac{w}{\mu - \Lambda \bar{F}(v_s)}.$$

The subscription threshold, v_s, is independent of r whereas $v_p(r)$ is a decreasing function of r - as third-party revenue increases the firm naturally charges a lower price to use the service. Given that the per-use price is

$$p_p = v_p - \frac{w}{\mu - \Lambda \bar{F}(v_p)},$$

it is possible, for r large enough, that the optimal per-use price pays customers for the use of the service ($p_p < 0$): if the firm earns sufficient revenue per transaction, then the firm has an incentive to encourage transactions.[16]

Differentiation yields,

$$\begin{aligned} \frac{\partial R_s(v_s; r)}{\partial r} &= \Lambda \bar{F}(v_s) \\ \frac{\partial R_p(v_p; r)}{\partial r} &= \Lambda \bar{F}(v_p(r)) \end{aligned}$$

From $v_s < v_p(0)$ and $dv_p(r)/dr < 0$, it follows that subscription revenue initially increases faster in r than per-use revenue but for some sufficiently high r, per-use revenue increases faster in r. We can establish that $v_s = v_p(\widetilde{r})$ for

$$\widetilde{r} = \frac{\bar{F}(v_s)}{f(v_s)} + \frac{w\Lambda \bar{F}(v_s)}{\left(\mu - \Lambda \bar{F}(v_s)\right)^2}.$$

[16] We have assumed the reward per transaction, r, is independent of the number of transactions. In practice, it would be a declining function of the number of transactions. Hence, we do not want to claim firms should pay customers to generate transactions (because they may generate transactions that cannot be converted into revenue). Instead, we are merely claiming that the presence of third-party transaction revenue may induce a firm to charge a low per-use price, possibly even zero.

When $r = \widetilde{r}$, the per-use price is zero, $p_p = 0$, and $R_p\left(v_p(\widetilde{r})\right) < R_s(v_s)$. Thus, if third-party revenue is such that the optimal per-use price is positive (i.e., the firm charges for the service rather than pays customers to use the service), then subscription profit increases faster in r than per-use profit. In this sense, third-party revenue favors subscription pricing over per-use pricing.[17]

All the results in this section hold for the industry standard model (section 4). Most results generalize for the endogenous capacity model as well (section 5). More specifically, in this case, we get that $\partial\Pi_s(v_s;r)/\partial r = \Lambda\bar{F}(v_s)$, $\partial\Pi_p(v_p;r)/\partial r = \Lambda\bar{F}(v_p(r))$, and $dv_p(r)/dr < 0$, which lead to favoring subscription over per-use pricing for $r < \widetilde{r}$. However, as opposed to the other two models, $v_s = v_p(\widetilde{r})$ does not occur at a zero per-use price.

The two-sided markets literature also finds that a firm may want to charge a low per-use price in a market. By doing so the firm creates a positive externality in a second market, thereby allowing the firm to increase the price it charges in that second market. However, our results demonstrate that the firm need not sacrifice revenue in one market to earn it in another - with subscription pricing the firm generates plenty of transactions (because the per-use fee is zero) while at the same time the firm collects revenues (from subscriptions). Thus, we conclude that the attractiveness of subscription pricing relative to per-use pricing is enhanced in the presence of third-party revenue.[18]

7 Conclusion

Using a queueing framework, we find that a firm may prefer subscription pricing over per-use pricing even if consumers dislike congestion. Furthermore, subscription pricing may be preferable in situations that would *a priori* suggest a preference for per-use pricing: when the industry has a standard that customers spend little time in the service process; or when customers' strongly dislike the time to complete the service thereby making congestion costly to the firm. Subscription pricing can dominate in these situations because (i) the firm must invest in a considerable amount of capacity to meet the standard or to reduce service times to a minimum and (ii) the firm can cover that large capacity cost only if it can extract enough revenue from customers. Next, we find that the absolute advantage of subscription pricing can be considerable whereas the absolute advantage of per-use pricing is generally modest - per-use pricing generates higher revenue or earns

[17] Note that by "favor" we do not mean $R_s(v_s) > R_p\left(v_p(r)\right), \forall r$. As we have previously indicated, there exist parameter values for which $R_p\left(v_p(0)\right) > R_s(v_s)$, which implies that there exists a range of rewards $r \in \{[0, r'] : r' < \widetilde{r}\}$ for which per-use revenue is higher than subscription revenue. Instead, we use "favor" to mean that $R_s(v_s) - R_p\left(v_p(r)\right)$ increases in r as long as $r < \widetilde{r}$. That is, if we restirct attention to non-negtive per-use prices, once subscription revenue is greater than per-use revenue, i.e. as r increases above r', it remains greater for all greater r.

[18] If $r < 0$, the the firm incurs a cost per transaction. In this case, an increase in the transaction cost (r decreases) intuitively favors per-use pricing over subscription pricing.

higher profit only when revenue or profit is reasonably low. Furthermore, some services are able to earn third-party revenue that is proportional to the usage of its service. As that revenue potential increases, subscription pricing becomes more attractive relative to per-use pricing - subscription pricing encourages transactions (which increases third-party revenue) at the same time that it earns revenue from customers. Overall, we conclude that the emphasis on per-use pricing in the queueing literature is misplaced - we provide evidence that subscription pricing can indeed be the preferable pricing strategy even in services that experience congestion.

Acknowledgments

The authors thank Albert Ha and the seminar participants at Hong Kong University of Science and Technology for their helpful feedback.

References

[1] Armstrong, M. 2006. Competition in two-sided markets. *The RAND Journal of Economics* **37**(3) 668-691.

[2] Auchard, E. 2007. EBay to resume ads on Google, but rely on rivals. *Reuters* (June 22).

[3] Barro, R.J., P.M. Romer. 1987. Ski-lift pricing, with applications to labor and other markets. *The American Economic Review* **77**(5) 875-890.

[4] Berglas, E. 1976. On the theory of clubs. *The American Economic Review* **66**(2) 116-121.

[5] Bitran, G., P. Rocha e Oliveira, A. Schilkrut. 2008. Managing customer relationships through pricing and service quality. Working Paper, MIT.

[6] Chen, H., M. Frank. 2004. Monopoly pricing when customers queue. *IIE Transactions* **36** 569-581.

[7] Cleveland, B., J. Mayben. 1997. *Call center management on fast forward*. Call Center Press, Annapolis, Maryland.

[8] DeGraba, P. 1995. Buying frenzies and seller-induced excess demand. *The RAND Journal of Economics* **26**(2) 331-342.

[9] De Vany, A. 1976. Uncertainty, waiting time, and capacity utilization: a stochastic theory of product quality. *Journal of Political Economy* **84**, 523-540.

[10] Edelson, M.M., D.K. Hilderbrand. 1975. Congestion tolls for Poisson queueing processes. *Econometrica* **43**(1) 81-92.

[11] Essegaier, S., S. Gupta, Z.J. Zhang. 2002. Pricing access services. *Marketing Science* **21**(2) 139-159.

[12] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: tutorial, review, and research prospects. *Manufacturing and Service Operations Management* **5**(2) 79-141.

[13] Hassin, R., M. Haviv. 2003. *To queue or not to queue: equilibrium behavior in queueing systems.* Kluwer Academic Publishers, Boston, Massachusetts.

[14] Littlechild, S.C. 1974. Optimal arrival rate in a simple queueing system. *International Journal of Production Research* **12**(3) 391-397.

[15] Mendelson, H. 1985. Pricing computer services: queueing effects. *Communications of the ACM* **28** 312-321.

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

[17] Randhawa, R., S. Kumar. Usage restriction and subscription services: operational benefits with rational customers, forthcoming in *Manufacturing and Service Operations Management*, 2008.

[18] Rochet, J-C., J. Tirole. 2006. Two-sided market: a progress report. *The RAND Journal of Economics* **37**(3) 645-667.

[19] Scotchmer, S. 1985. Profit-maximizing clubs. *Journal of Public Economics* 27(1) 25-45.

[20] Xie, J., S.M. Shugan. 2001. Electronic tickets, smart cards, and online prepayments: when and how to advance sell. *Marketing Science* **20**(3) 219–243.

Appendix A: Proofs

Theorems 1 and 4 hold if F exhibits an increasing generalized failure rate (IGFR). F is IGFR if and only if $xh(x)$ is increasing, where $h(x) = f(x)/\bar{F}(x)$ is the failure rate. That is, the IGFR property is more general than the IFR.

Proof of Theorem 1. That (2) is a local maximum is shown by examining the first-order conditions of $R_p(v)$.

$$\frac{dR_p(v)}{dv} = -\Lambda f(v)\left(v - wW(\Lambda\bar{F}(v))\right) + \Lambda\bar{F}(v)\left(1 + w\Lambda f(v)W'\left(\Lambda\bar{F}(v)\right)\right)$$

Let $\varrho(v) = dR_p(v)/dv$. Noting that $\varrho(0) > \Lambda$ and that $\lim_{v\to\infty}\varrho(v) \leq 0$, indicates that there exists at least one maximum. Showing that the solution v_p is unique will complete the proof. v_p satisfies the first-order condition given by (2). Rearranging the terms in (2) gives:

$$1 - \frac{\bar{F}(v_p)}{v_p f\left(v_p\right)} = \frac{wW(\Lambda\bar{F}(v_p))}{v_p} + \frac{w\Lambda\bar{F}(v_p)W'(\Lambda\bar{F}(v_p))}{v_p}$$

Since F is IGFR, $f\left(v_p\right)v_p/\bar{F}(v_p)$ is weakly increasing in v_p, and thus the term in the LHS is also increasing in v_p. The term in the RHS is strictly decreasing in v_p, because $\bar{F}(v) = \Pr\{V \geq v\}$ is decreasing in v and $W(\cdot)$ is increasing and convex. ■

Proof of Lemma 1. For the per-use case, we have established in Theorem 1 that $R_p(v)$ is quasi-concave. Also note that $R_p(\overline{v}) = 0$. $\overline{v}$ will be the maximizer, when the first-order conditions at $\overline{v}$ are non negative. Evaluating the first-order conditions at $v_p = \overline{v}$ we get: $\Lambda f(\overline{v})\left(w/\mu - \overline{v}\right) \geq 0$, which is equivalent to $w/\Lambda \geq \overline{v}/\rho$. Since $R_p(\overline{v}) = 0$, this will imply that $R_p(v) < 0\ \forall v \in [0, \overline{v})$. Thus, we can limit our search space to the interesting case, $w/\Lambda < \overline{v}/\rho$. In this case there exists an interior solution that results in a positive per-use revenue rate. Performing a similar analysis for the subscription case reveals that the condition for an interior maximum that guarantees positive revenue rates is the same as in the per-use case, namely we must have $w/\Lambda < \overline{v}/\rho$.

Next, we show that for any given α and ρ, the revenue functions are linear in $\overline{v}$, which implies that $\overline{v}$ does not affect the comparison between subscription and pay per use. The comparison can be made solely on the basis of α and ρ. To see that this is indeed the case, note that $v_s/\overline{v}$ can be implicitly expressed by:

$$\frac{v_s}{\overline{v}} = \frac{\alpha}{1 - \rho\left(1 - \frac{v_s}{\overline{v}}\right)} \tag{10}$$

Thus, $v_s = \overline{v} \cdot g\left(\alpha, \rho\right)$, where g is a function of α and ρ only. Plugging v_s into the subscription revenue function, we obtain:

$$R_s = \frac{\Lambda\left(1 - g\left(\alpha, \rho\right)\right)^2}{2} \cdot \overline{v}$$

which is linear in $\overline{v}$. Similarly for the pay-per-use case, we can express $v_p/\overline{v}$ by:

$$\frac{v_p}{\overline{v}} = \frac{1}{2}\left(1 + \frac{\alpha}{\left(1 - \rho\left(1 - \frac{v_p}{\overline{v}}\right)\right)^2}\right) \tag{11}$$

Thus, $v_p = \overline{v} \cdot h(\alpha, \rho)$, where h is a function of α and ρ only. Plugging v_p into the subscription revenue function, we obtain:

$$R_p = \Lambda (1 - h(\alpha, \rho)) \left(h(\alpha, \rho) - \frac{\alpha}{1 - \rho (1 - h(\alpha, \rho))} \right) \cdot \overline{v}$$

which is also linear in $\overline{v}$. Thus, to compare the revenues of the two schemes it suffices to examine changes in α and ρ. ■

Proof of Theorem 2. Uniqueness: note first, that $\lim_{\rho \to 0} R_s(\rho) = \frac{(1-\alpha)^2}{2}$ and $\lim_{\rho \to 0} R_p(\rho) = \frac{(1-\alpha)^2}{4}$. Also, $\lim_{\rho \to \infty} R_s(\rho) = \lim_{\rho \to \infty} R_p(\rho) = 0$. Finally, we show that both R_s and R_p are monotonically decreasing in ρ and that $R_s'(\rho) < R_p'(\rho)\ \forall \rho$. By implicitly differentiating the revenue functions, we get:

$$\frac{dR_s(\rho)}{d\rho} = -\frac{\alpha (1 - g)^2}{(1 - \rho (1 - g))^2}$$

and

$$\frac{dR_p(\rho)}{d\rho} = -\frac{\alpha (1 - h)^2}{(1 - \rho (1 - h))^2}$$

where g and h are shorthand notation for $g(\alpha, \rho)$ and $h(\alpha, \rho)$. That $R_s'(\rho) < R_p'(\rho)\ \forall \rho$ follows from the fact that $v_s(\rho) < v_p(\rho)\ \forall \rho$ (see online companion for proof).

The threshold utilization factor, $\widetilde{\rho}$, solves:

$$F = \frac{(1 - g(\alpha, \widetilde{\rho}))^2}{2} - (1 - h(\alpha, \widetilde{\rho})) \left(h(\alpha, \widetilde{\rho}) - \frac{\alpha}{1 - \widetilde{\rho} (1 - h(\alpha, \widetilde{\rho}))} \right) = 0 \tag{12}$$

Implicitly differentiating (12) with respect to α yields:

$$\frac{dF}{d\alpha} = \left(1 - \frac{v_s}{\overline{v}}\right) \frac{dg}{d\alpha} + \frac{1 - \frac{v_p}{\overline{v}}}{1 - \widetilde{\rho} \left(1 - \frac{v_p}{\overline{v}}\right)}$$

which is positive because v_s is increasing in α. Similarly, implicit differentiation with respect to $\widetilde{\rho}$ yields:

$$\frac{dF}{d\widetilde{\rho}} = \left(1 - \frac{v_s}{\overline{v}}\right) \frac{dg}{d\widetilde{\rho}} + \frac{\alpha \left(1 - \frac{v_p}{\overline{v}}\right)^2}{\left(1 - \widetilde{\rho} \left(1 - \frac{v_p}{\overline{v}}\right)\right)^2}$$

which is positive because v_s is increasing in $\widetilde{\rho}$. Applying the implicit function theorem we then get that $d\widetilde{\rho}(\alpha)/d\alpha < 0$ ■

Proof of Theorem 3. *(i)* $\rho = 0$: Substituting $\rho = 0$ in equations (10) and (11) results in $v_s = \alpha \overline{v}$ and $v_p = (1 + \alpha)\overline{v}/2$. Then, the following expressions for the revenue rates are immediate:

$$R_s = \frac{\Lambda \overline{v} (1 + \alpha)^2}{2};\ R_p = \frac{\Lambda \overline{v} (1 + \alpha)^2}{4}$$

(ii) $\rho \to \infty$: Rearranging (10), we get:

$$\frac{v_s}{\overline{v}}\left(\frac{1-\rho}{\rho}+\frac{v_s}{\overline{v}}\right)=\frac{\alpha}{\rho}$$

As $\rho \to \infty$, there are up to two roots that solve the above. The larger of the two is a maximum. This implies that in this case, we have

$$\lim_{\rho\to\infty}\frac{v_s}{\overline{v}}=\frac{\rho-1}{\rho}.$$

Similarly, rewriting (11), we get

$$\left(\frac{v_p}{\overline{v}}-\frac{1}{2}\right)\left(\frac{1-\rho}{\rho}+\frac{v_p}{\overline{v}}\right)^2=\frac{\alpha}{2\rho^2}.$$

This implies that for all pricing schemes,

$$\lim_{\rho\to\infty}\frac{v_s}{\overline{v}}=\lim_{\rho\to\infty}\frac{v_p}{\overline{v}}=\frac{\rho-1}{\rho}$$

Substituting into the revenue rates, we obtain for $\rho \to \infty$:

$$R_s=\frac{\Lambda\overline{v}}{2\rho^2};\ R_p=\frac{\Lambda\overline{v}\,(\rho-1)}{\rho^2}.$$

Thus, it follows that $\lim_{\rho\to\infty} R_x = 0$, $x \in \{s, p\}$ and that

$$\lim_{\rho\to\infty}\frac{R_s}{R_p}=\lim_{\rho\to\infty}\frac{1}{2\rho-2}=0.$$

■

Proof of Theorem 4. That (5) is a local maximum is shown by examining the first-order conditions of $\Pi_p(v)$.

$$\frac{d\Pi_p(v)}{dv}=-\Lambda f(v)\left(v-\frac{w}{I}-c\right)+\Lambda\bar{F}(v)$$

Let $\varrho(v) = d\Pi_p(v)/dv$. Noting that $\varrho(0) > \Lambda$ and that $\lim_{v\to\infty}\varrho(v) \leq 0$, indicates that there exists at least one maximum. Showing that the solution v_p is unique will complete the proof. v_p satisfies the first-order condition given by (5). Rearranging the terms in (5) gives:

$$1-\frac{\bar{F}(v_p)}{v_p f\left(v_p\right)}=\frac{w}{Iv_p}+\frac{c}{v_p}$$

Since F is IGFR, $f\left(v_p\right)v_p/\bar{F}(v_p)$ is weakly increasing in v_p, and thus the term in the LHS is also increasing in v_p. The term in the RHS is strictly decreasing in v_p. ■

Proof of Theorem 5. Substituting $v_s = w/I$ and $v_p = \frac{\overline{v}+w/I+c}{2}$ in $\Pi_s(I)$ and $\Pi_p(I)$, respectively:

$$\Pi_s(I) = \Lambda\left(\frac{\overline{v}-w/I}{\overline{v}}\right)\left(\frac{\overline{v}-w/I-2c}{2}\right) - cI$$

and

$$\Pi_p(I) = \frac{\Lambda(\overline{v}-w/I-c)^2}{4\overline{v}} - cI$$

Differentiating with respect to I, we get:

$$\frac{\partial \Pi_s}{\partial I} = \frac{\Lambda w}{I^2}\left(\frac{\overline{v}-w/I-c}{\overline{v}}\right) - c$$

and

$$\frac{\partial \Pi_p}{\partial I} = \frac{\Lambda w}{I^2}\left(\frac{\overline{v}-w/I-c}{2\overline{v}}\right) - c.$$

Comparing both equations, we get the desired result. ■

Proof of Theorem 6. We prove by contradiction that if there is a maximum, it is unique. Suppose there exist v_1 and v_3 such that $v_1 < v_3$, $\Pi_p(v_1) \geq 0$, $\Pi_p(v_3) \geq 0$, $\Pi_p'(v_1) = 0$, $\Pi_p'(v_3) = 0$, i.e., both are local maxima with positive profits. Our conditions imply for both $v \in \{v_1, v_3\}$ that

$$\begin{aligned} v - c &= \frac{\bar{F}(v)}{f(v)} + \frac{\phi}{\sqrt{\bar{F}(v)}} \\ \frac{v-c}{2} &\geq \frac{\phi}{\sqrt{\bar{F}(v)}} \end{aligned}$$

Combining the two conditions we have

$$\frac{\bar{F}(v)}{f(v)} \geq \frac{\phi}{\sqrt{\bar{F}(v)}}$$

Given that v_1 and v_3 are local maxima, there must be a local minima, v_2, such that $v_1 < v_2 < v_3$. There are two cases to consider: $\Pi_p(v_2) < 0$ and $\Pi_p(v_2) > 0$.

Consider $\Pi_p(v_2) < 0$. Analogous to (6), $\Pi_p'(v_2) = 0$ and $\Pi_p(v_2) < 0$ imply

$$\frac{\bar{F}(v_2)}{f(v_2)} < \frac{\phi}{\sqrt{\bar{F}(v_2)}} \tag{13}$$

Because F is IFR, the left hand side of (13) is decreasing. Furthermore, the right hand side of (13) is increasing. As a result, $v_1 < v_2 < v_3$ implies

$$\frac{\bar{F}(v_1)}{f(v_1)} > \frac{\bar{F}(v_2)}{f(v_2)} > \frac{\bar{F}(v_3)}{f(v_3)} \tag{14}$$

and

$$\frac{\phi}{\sqrt{\bar{F}(v_3)}} > \frac{\phi}{\sqrt{\bar{F}(v_2)}} > \frac{\phi}{\sqrt{\bar{F}(v_1)}}. \tag{15}$$

Combining (13), (14) and (15) yields

$$\frac{\bar{F}(v_3)}{f(v_3)} < \frac{\bar{F}(v_2)}{f(v_2)} < \frac{\phi}{\sqrt{\bar{F}(v_2)}} < \frac{\phi}{\sqrt{\bar{F}(v_3)}},$$

which contradicts (6).

Consider the second case, $\Pi_p(v_2) > 0$. Denote

$$z(v) = -v + \frac{\bar{F}(v)}{f(v)} + \frac{\phi}{\sqrt{\bar{F}(v)}} + c$$

Differentiate:

$$z'(v) = -1 - \left(\frac{f'(v)\bar{F}(v) + f(v)^2}{f(v)^2}\right) + \frac{f(v)}{2\bar{F}(v)}\frac{\phi}{\sqrt{\bar{F}(v)}}$$

Given that F is IFR, the second term is negative. (6) implies the third term is less than 1/2. Hence, $z'(v) < 0$ for v_1, v_2 and v_3. Because $\Pi_p'(v_1) = \Pi_p'(v_2) = \Pi_p'(v_3) = 0$, it follows that $z(v_1) = z(v_2) = z(v_2) = 0$. However, due to the continuity of $z(v)$, this is not feasible if $z'(v) < 0$ for v_1, v_2 and v_3.

Observe that $\Pi_p(0) = -\Lambda c - 2\sqrt{cw\Lambda}$ is negative and that $\lim_{v\to\bar{v}} \Pi_p(v) = 0$. Given that $\Pi_p(0)$ is finite and $\lim_{v\to\bar{v}} \Pi_p(v) = 0$, a maximum exists if there exists a $v_p < \bar{v}$ such that $\Pi_p'(v_p) = 0$ and $\Pi_p(v_p) \geq 0$. Requiring that $\Pi_p(v_p) \geq 0$ is equivalent to having

$$\frac{\Pi_p(v)}{\Lambda\bar{F}(v)} = v - c - \frac{2\phi}{\sqrt{\bar{F}(v)}} \geq 0$$

for some v**.** Assume $\phi = 0$. If $\bar{v} > c$, there must be a solution with positive profit. Let $M_p(\phi) \equiv \Pi_p(v_p(\phi), \phi)$. From the Envelope Theorem, we have:

$$\frac{\partial M_p(\phi)}{\partial \phi} = -\Lambda\sqrt{\bar{F}(v_p)} < 0$$

which means that $\Pi_p(v_p(\phi), \phi)$ is decreasing in ϕ. This implies there exists some $\bar{\phi}_p$ such that $\Pi_p(v_p(\phi), \phi)$ for $\phi \leq \bar{\phi}_p$. Otherwise, there does not exist an optimal $v_p < \bar{v}$.

v_p is the smallest solution to (7): While we cannot show the number of possible solutions to (7) for a general IFR distribution F, we show by contradiction, that the optimal v_p is the smallest solution to (7). Suppose there exist v_1 and v_3 such that $v_1 < v_3$, $\Pi_p(v_1) < 0$, $\Pi_p(v_3) < 0$, $\Pi_p'(v_1) = 0$, $\Pi_p'(v_3) = 0$, i.e., both are local maxima with negative profits. Our conditions imply for

both $v \in \{v_1, v_3\}$ that

$$\begin{aligned} v - c &= \frac{\bar{F}(v)}{f(v)} + \frac{\phi}{\sqrt{\bar{F}(v)}} \\ \frac{v-c}{2} &< \frac{\phi}{\sqrt{\bar{F}(v)}} \end{aligned}$$

Combining the two conditions we have

$$\frac{\bar{F}(v)}{f(v)} < \frac{\phi}{\sqrt{\bar{F}(v)}}$$

Assume there exists a local maxima, v_2, such that $v_1 < v_2 < v_3$ and $\Pi_p(v_2) > 0$. This implies that

$$\frac{\bar{F}(v_2)}{f(v_2)} > \frac{\phi}{\sqrt{\bar{F}(v_2)}} \tag{16}$$

Consider v_1. Because F is IFR, $v_1 < v_2$ implies:

$$\frac{\bar{F}(v_1)}{f(v_1)} > \frac{\bar{F}(v_2)}{f(v_2)} \tag{17}$$

and

$$\frac{\phi}{\sqrt{\bar{F}(v_2)}} > \frac{\phi}{\sqrt{\bar{F}(v_1)}}. \tag{18}$$

Combining conditions (16) and (18), we get:

$$\frac{\bar{F}(v_2)}{f(v_2)} > \frac{\phi}{\sqrt{\bar{F}(v_2)}} > \frac{\phi}{\sqrt{\bar{F}(v_1)}} > \frac{\bar{F}(v_1)}{f(v_1)}$$

which contradicts condition (17). Letting $v_2 < v_3$, however, a contradiction cannot be reached, which does not preclude the existence of additional solutions to (7) in the negative range for a general IFR distribution.

Sufficient condition for at most two solutions to (7): Rearranging (7), we have:

$$v_p - \frac{1}{h(v_p)} = \frac{\phi}{\sqrt{\bar{F}(v_p)}} + c.$$

The RHS is convex and increasing and the LHS is increasing. Taking the derivative of the LHS, we get $1 + h'(v_p)/\left(h(v_p)\right)^2$. Thus, if condition (A1) holds, there can be at most two solutions to (7), with the smallest one being the maximum. ■

Proof of Theorem 7. First note that $\Pi_s(0) = -\infty$ and that $\lim_{v \to \overline{v}} \Pi_s(v) = 0$. Differentiating $\Pi_s(v)$, we obtain:

$$\frac{d\Pi_s(v)}{dv} = c\left(\frac{w}{v^2} + \Lambda f(v)\right) - \Lambda \bar{F}(v)$$

Equating to zero and rearranging terms, the result in (9) follows.

A maximum exists if there exists a $v_s < \bar{v}$ such that $\Pi_s'(v_s) = 0$ and $\Pi_s(v_s) \geq 0$. Requiring that $\Pi_s(v_s) \geq 0$ is equivalent to having

$$\frac{\Pi_s(v)}{\Lambda \bar{F}(v)} = E\left[V|V \geq v\right] - v - c - \frac{\phi^2}{v\bar{F}(v)} \geq 0$$

for some v. Assume $\phi = 0$. Then, if $E\left[V\right] > c$, there must be a solution with positive profit. Let $M_s(\phi) \equiv \Pi_s(v_s\left(\phi\right), \phi)$. From the Envelope Theorem, we have:

$$\frac{\partial M_s(\phi)}{\partial \phi} = -\frac{2\phi\Lambda}{v_s} < 0$$

which means that $\Pi_s(v_s\left(\phi\right), \phi)$ is decreasing in ϕ. Note that even though we have not ruled out the existence of several local maxima v_s, $\Pi_s(v_s\left(\phi\right), \phi)$ is decreasing in ϕ at every critical point. This implies there exists some $\overline{\phi}_s$ such that $\Pi_s(v_s\left(\phi\right), \phi)$ for $\phi \leq \overline{\phi}_s$. Otherwise, there does not exist an optimal $v_s < \bar{v}$.

Furthermore, let

$$z(v) = \phi\left(1 - c \cdot h\left(v\right)\right)^{-\frac{1}{2}} \left(\bar{F}(v)\right)^{-\frac{1}{2}}.$$

We want to show that there exists at most one v_s that maximizes profit and solves $v_s = z(v_s)$. Differentiating $z(v)$, we get:

$$z'(v) = \frac{\phi}{2}\left(c \cdot h'\left(v\right)\left(1 - c \cdot h\left(v\right)\right)^{-\frac{3}{2}} \left(\bar{F}(v)\right)^{-\frac{1}{2}} + f\left(v\right)\left(1 - c \cdot h\left(v\right)\right)^{-\frac{1}{2}} \left(\bar{F}(v)\right)^{-\frac{3}{2}}\right)$$

Plugging in (9), we get:

$$z'(v_s) = \frac{\phi}{2}\left(h\left(v_s\right)v_s + \frac{c \cdot h'\left(v_s\right)v_s}{1 - c \cdot h\left(v_s\right)}\right).$$

A sufficient condition for $z'(v_s)$ to be increasing is for both terms in the brackets to be increasing. The first term is the generalized failure rate. It is increasing if F is IGFR. The second term is increasing if $h'\left(v_s\right)v_s$ is increasing and F is IFR. ■

Technical Appendix to "Pricing Services Subject to Congestion: Charge Per-Use Fees or Sell Subscriptions?"

Gérard P. Cachon • Pnina Feldman

Operations and Information Management, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA

cachon@wharton.upenn.edu • pninaf@wharton.upenn.edu

In this technical appendix we provide the proofs of additional results relating to the main text of the paper. Throughout this appendix, we introduce propositions that are not explicitly included in the main text of the paper but are useful for deriving some of the intuitions and results that we claim in the main text without proof. In addition, we report on a numerical study that generalizes the analytical results. The first section includes the proofs from the exogenous capacity models (§4 in the main text), while the second section provides the proofs from the industry time standards section (§5 in the main text) and the third section provides the proofs of the endogenous capacity case discussed in §6 in the main text. The fourth section generalizes the fixed capacity model to linear waiting costs and the fifth section analyses the models to include mixed strategy equilibria. The sixth section reports on a numerical study performed to compare between the pricing schemes in the endogenous capacity case and the seventh section generalizes the third-party revenue result for strictly concave and increasing revenue functions.

1 Proofs from the Exogenous Capacity Model

For the comparison between per-use and subscription, it is useful to introduce a third pricing scheme–the two-part tariff. The two-part tariff, combines a per-use fee and a subscription price: e.g., a customer may have to pay a fee to join a golf club and then a fee for each round of golf actually played. Where useful, we use the "t" subscript to signify notation associated with the two-part tariff scheme.

A two-part tariff combines a per-use fee, p, with a subscription rate. As with per-use pricing, a consumer with a value V for a service opportunity is indifferent between seeking service or not when $V \geq p + wW(\lambda)$. Therefore, the arrival rate to the firm is $\Lambda\bar{F}(v)$, where v is the unique solution to $v = p + wW(\Lambda\bar{F}(v))$. Revenue from per-use fees accrues at rate $R_p(v)$, just as in the per-use scheme, assuming all consumers are subscribers. As in the subscription case, the optimal subscription makes all M consumers indifferent between purchasing the subscription or not. Hence, subscription revenues are $R_s(v)$. Total revenue from the two-part tariff is then expressed in terms of the threshold for the subscription customer who is indifferent between paying the per-use fee for service or not:

$$\begin{aligned} R_t(v) &= R_p(v) + R_s(v) \\ &= \Lambda\bar{F}(v)\left(E\left[V|V \geq v\right] - wW(\Lambda\bar{F}(v))\right) \end{aligned}$$

Note, unlike with subscription pricing, and just like with per-use pricing, the threshold v is a decision variable for the firm. The following proposition establishes that there is a unique optimal threshold for the firm.

Proposition 1 *The firm's revenue function with a two-part tariff, $R_t(v)$, is quasi-concave. Hence, there exists a unique optimal threshold, $v_t = \arg\max_v R_t(v)$, where v_t is the unique solution to*

$$v_t = wW(\Lambda\bar{F}(v_t)) + w\Lambda\bar{F}(v_t)W'(\Lambda\bar{F}(v_t)) \tag{1}$$

Proof. *Existence:* To prove that the function $R_t(v)$ has at least one local maximum, write the first-order conditions:

$$\frac{dR_t(v)}{dv} = \Lambda f(v)\left(wW\left(\Lambda\bar{F}(v)\right) + w\Lambda\bar{F}(v)W'\left(\Lambda\bar{F}(v)\right) - v\right)$$

Let $\varrho(v) = dR_t(v)/dv$ and examine $\varrho(v)$ at the limit points. Since $\varrho(0) > 0$ and $\lim_{v\to\infty}\varrho(v) \leq 0$ and hence is less than $\varrho(0)$, it follows that $R_t(v)$ has at least one maximum in the range. *Uniqueness:* solving $R_t(v) = 0$, we get the FOC given in (1). Since the LHS is strictly increasing and the RHS is strictly decreasing, $R_t(v)$ is quasi-concave and there exists a unique threshold, v_t, that is implicitly defined by condition (1). ■

Proposition 2 *The resulting two-part tariff threshold valuation, above which a customer requests service, lies in between the threshold valuation in the pay-per-use case and that of the subscription case, i.e., $v_s < v_t < v_p$.*

Proof. Denote right-hand-sides functions of conditions

$$v_p = \frac{\bar{F}(v_p)}{f(v_p)} + wW(\Lambda\bar{F}(v_p)) + w\Lambda\bar{F}(v_p)W'(\Lambda\bar{F}(v_p)), \tag{2}$$

$v_s = wW\left(\Lambda\bar{F}(v_s)\right)$ and (1) by $g_p(v)$, $g_s(v)$ and $g_t(v)$, respectively. Then, $g_s(v) < g_t(v) < g_p(v)\ \forall v$. Since the LHS is strictly increasing in v, the result follows. ■

The following statement of Lemma 1 is more general than the statement of Lemma 1 in the main text because it includes results about the two-part tariff pricing scheme.

Lemma 1 *The optimal revenue with each pricing scheme (R_p, R_s and R_t) can be expressed in terms of α, ρ, Λ and $\overline{v}$, where $\alpha = w/\mu\overline{v}$ and $\rho = \Lambda/\mu$. Furthermore, all three revenues are non-negative for $\alpha \in [0,1]$ and linear in $\Lambda\overline{v}$. Consequently, the relative revenues (e.g., R_p/R_t, R_s/R_t, and R_p/R_s) can be expressed in terms of α and ρ.*

Proof. For the social welfare case, we have established in Proposition 1 that $R_t(v)$ is quasi-concave. Also note that $R_t(\overline{v}) = 0$. $\overline{v}$ will be the maximizer, when the first-order conditions at $\overline{v}$ are non negative. Evaluating the first-order conditions at $v_t = \overline{v}$ we get:

$$\frac{\Lambda}{\overline{v}}\left(\frac{w\mu}{\left(\mu - \Lambda\left(\frac{\overline{v}-v_t}{\overline{v}}\right)\right)^2} - v_t\right)\Bigg|_{v_t=\overline{v}} = \frac{\Lambda}{\overline{v}}\left(\frac{w}{\mu} - \overline{v}\right) \geq 0$$

which is equivalent to $w/\Lambda \geq \overline{v}/\rho$. Since $R_t(\overline{v}) = 0$, this will imply that $R_t(v) < 0\ \forall v \in [0,\overline{v})$. Since the revenue rates in the other two schemes are never higher than in the social optimal case, we can limit our search space to the interesting case, $w/\Lambda < \overline{v}/\rho$. In this case there exists an interior solution that results in a positive social optimal revenue rate. Performing a similar analysis for the pay-per-use and subscription cases, reveals that the conditions for an interior maximum that guarantees positive revenue rates are the same as in the social optimal case, namely we must have $w/\Lambda < \overline{v}/\rho$.

Next, we show that for any given α and ρ, the revenue functions are linear in $\overline{v}$, which implies that $\overline{v}$ does not affect the comparison between subscription and pay per use. The comparison can be made solely on the basis of α and ρ. To see that this is indeed the case, note that $v_s/\overline{v}$ can be implicitly expressed by:

$$\frac{v_s}{\overline{v}} = \frac{\alpha}{1-\rho\left(1-\frac{v_s}{\overline{v}}\right)} \tag{3}$$

Thus, $v_s = \overline{v} \cdot g(\alpha, \rho)$, where g is a function of α and ρ only. Plugging v_s into the subscription revenue function, we obtain:

$$R_s = \frac{\Lambda\left(1-g(\alpha,\rho)\right)^2}{2} \cdot \overline{v}$$

which is linear in $\overline{v}$. Similarly for the pay-per-use case, we can express $v_p/\overline{v}$ by:

$$\frac{v_p}{\overline{v}} = \frac{1}{2}\left(1+\frac{\alpha}{\left(1-\rho\left(1-\frac{v_p}{\overline{v}}\right)\right)^2}\right) \tag{4}$$

Thus, $v_p = \overline{v} \cdot h(\alpha, \rho)$, where h is a function of α and ρ only. Plugging v_p into the subscription revenue function, we obtain:

$$R_p = \Lambda\left(1-h(\alpha,\rho)\right)\left(h(\alpha,\rho) - \frac{\alpha}{1-\rho\left(1-h(\alpha,\rho)\right)}\right) \cdot \overline{v}$$

which is also linear in $\overline{v}$. Thus, to compare the revenues of the two schemes it suffices to examine changes in α and ρ.

Finally, for the social optimal case,

$$\frac{v_t}{\overline{v}} = \frac{\alpha}{\left(1-\rho\left(1-\frac{v_t}{\overline{v}}\right)\right)^2} \tag{5}$$

Thus, $v_t = \overline{v} \cdot l(\alpha, \rho)$, where l is a function of α and ρ only. Plugging v_t into the subscription revenue function, we obtain:

$$R_t = \Lambda\left(1-l(\alpha,\rho)\right)\left(\frac{1+l(\alpha,\rho)}{2} - \frac{\alpha}{1-\rho\left(1-l(\alpha,\rho)\right)}\right) \cdot \overline{v}$$

which is again linear in $\overline{v}$. ■

Proposition 3 *When $\alpha = 0$, subscription is better than pay-per-use when $\rho < \sqrt{2}$ and vice versa. When $\alpha \to 1$, subscription is better than pay-per-use when $\rho < 1$.*

Proof. *(i) $\alpha = 0$. Pay-per-use.* Rearranging (4) we get:

$$\left(\frac{v_p}{\overline{v}} - \frac{1}{2}\right)\left(\frac{1-\rho}{\rho} + \frac{v_p}{\overline{v}}\right)^2 = \frac{\alpha}{2\rho^2} \tag{6}$$

If $\alpha = 0$, there are up to two roots that solve the above. The larger of the two is a maximum. This implies

$$\frac{v_p}{\overline{v}} = \begin{cases} \frac{1}{2} & \text{if } \rho \leq 2 \\ \frac{\rho-1}{\rho} & \text{if } \rho > 2 \end{cases}.$$

Substituting, it follows that

$$R_p(v_p;\alpha=0)=\begin{cases}\frac{\Lambda\overline{v}}{4} & \text{if } \rho\leq 2\\ \frac{\Lambda\overline{v}}{\rho^2}(\rho-1) & \text{if } \rho>2\end{cases}$$

Subscription. Rearranging (3), we get:

$$\frac{v_s}{\overline{v}}\left(1-\rho+\rho\frac{v_s}{\overline{v}}\right)=\alpha \tag{7}$$

When $\alpha=0$, there are up to two roots that solve the above. The larger of the two is a maximum. This implies that in this case, we have

$$\frac{v_s}{\overline{v}}=\begin{cases}0 & \text{if } \rho\leq 1\\ \frac{\rho-1}{\rho} & \text{if } \rho>1\end{cases}.$$

It thus follows that

$$R_s(v_s;\alpha=0)=\begin{cases}\frac{\Lambda\overline{v}}{2} & \text{if } \rho\leq 1\\ \frac{\Lambda\overline{v}}{2\rho^2} & \text{if } \rho>1\end{cases}$$

Comparing the two revenue functions, we get that $\widetilde{\rho}(0)=\sqrt{2}$. The corresponding revenue functions are then given by $R_s=R_p=\frac{\Lambda\overline{v}}{4}$.

(ii) $\alpha\to 1$. Rearranging (3) and substituting $\alpha=1$ we get the quadratic equation: $\rho\left(\frac{v_s}{\overline{v}}\right)^2+(1-\rho)\frac{v_s}{\overline{v}}-1=0$. The relevant root suggests $v_s\to\overline{v}$. As $v_p>v_s$ (Proposition 2), it follows that we must have $v_p\to\overline{v}$ as well. Next, we conjecture that $\widetilde{\rho}(1)=1$ and verify that this is indeed the case. Substituting $\rho=1$ in (6) and in (7), we get:

$$\left(\frac{v_p}{\overline{v}}-\frac{1}{2}\right)\left(\frac{v_p}{\overline{v}}\right)^2=\frac{\alpha}{2}$$

and

$$\left(\frac{v_s}{\overline{v}}\right)^2=\alpha,$$

respectively. Note that $\alpha=1$ and $v_s=v_p=\overline{v}$ solve both equations. Also note that this solution implies that $R_s=R_p$ and thus satisfies $\widetilde{\rho}(1)=1$. ■

The following statement of Theorem 1 is more general than the statement of Theorem 1 in the main text because it includes results about the two-part tariff pricing scheme. This version of Theorem 1 implies that subscription pricing approaches two-part tariff revenues when the potential utilization is low and that per-use pricing approaches two-part tariff for a high potential utilization.

Theorem 1 *The following limits hold: (i) When* $\rho\to 0$, $R_s=R_t=\Lambda\overline{v}(1+\alpha)^2/2$ *and* $R_p=\Lambda\overline{v}(1+\alpha)^2/4$. *(ii) When* $\rho\to\infty$, $\lim_{\rho\to\infty}R_p/R_t=1$ *while* $\lim_{\rho\to\infty}R_s/R_t=0$ *and* $\lim_{\rho\to\infty}R_x=0$, $x\in\{s,p,t\}$.

Proof. *(i)* $\rho=0$: Substituting $\rho=0$ in equations (3), (4) and (5) results in $v_s=v_t=\alpha\overline{v}$ and $v_p=(1+\alpha)\overline{v}/2$. Then, the following expressions for the revenue rates are immediate:

$$R_s=R_t=\frac{\Lambda\overline{v}(1+\alpha)^2}{2};\ R_p=\frac{\Lambda\overline{v}(1+\alpha)^2}{4}$$

(ii) $\rho \to \infty$: Rearranging (3), we get:

$$\frac{v_s}{\overline{v}}\left(\frac{1-\rho}{\rho}+\frac{v_s}{\overline{v}}\right)=\frac{\alpha}{\rho}$$

As $\rho \to \infty$, there are up to two roots that solve the above. The larger of the two is a maximum. This implies that in this case, we have

$$\lim_{\rho\to\infty}\frac{v_s}{\overline{v}}=\frac{\rho-1}{\rho}.$$

Similarly, rewriting (4) and (5), we get

$$\left(\frac{v_p}{\overline{v}}-\frac{1}{2}\right)\left(\frac{1-\rho}{\rho}+\frac{v_p}{\overline{v}}\right)^2=\frac{\alpha}{2\rho^2}$$

and

$$\frac{v_t}{\overline{v}}\left(\frac{1-\rho}{\rho}+\frac{v_t}{\overline{v}}\right)^2=\frac{\alpha}{\rho^2},$$

respectively. This implies that for all pricing schemes,

$$\lim_{\rho\to\infty}\frac{v_s}{\overline{v}}=\lim_{\rho\to\infty}\frac{v_p}{\overline{v}}=\lim_{\rho\to\infty}\frac{v_t}{\overline{v}}=\frac{\rho-1}{\rho}$$

Substituting into the revenue rates, we obtain for $\rho \to \infty$:

$$R_s=\frac{\Lambda\overline{v}}{2\rho^2};\ R_p=\frac{\Lambda\overline{v}\,(\rho-1)}{\rho^2};\ R_t=\frac{\Lambda\overline{v}\,(2\rho-1)}{2\rho^2}.$$

Thus, it follows that $\lim_{\rho\to\infty} R_x = 0$, $x \in \{s,p,t\}$ and that

$$\lim_{\rho\to\infty}\frac{R_s}{R_t}=\lim_{\rho\to\infty}\frac{1}{2\rho-1}=0$$

and

$$\lim_{\rho\to\infty}\frac{R_p}{R_t}=\lim_{\rho\to\infty}\frac{2\rho-2}{2\rho-1}=1$$

which indicates that per-use is a good approximation of the social optimal when $\rho \to \infty$. ■

Proposition 4 *The actual utilization rates given by* $\rho\overline{F}\left(v_x\left(\rho\right)\right)$, $x \in \{s,p,t\}$ *are strictly increasing in* ρ.

Proof. *Two-part-tariff:* Applying the Implicit Function Theorem to (5), we get:

$$\frac{\partial v_t}{\partial \rho}=\frac{y}{1+y\frac{\rho}{\overline{v}}}\cdot\frac{\overline{v}-v_t}{\overline{v}}$$

where y is given by

$$y=\frac{2\alpha\overline{v}}{\left(1-\rho\left(\frac{\overline{v}-v_t}{\overline{v}}\right)\right)^3}$$

Differentiating the utilization rate with respect to ρ leads to:

$$\begin{aligned}\frac{\partial \rho \overline{F}\left(v_t\left(\rho\right)\right)}{\partial \rho} &= \overline{F}\left(v_t\left(\rho\right)\right) - \rho \cdot f\left(v_t\left(\rho\right)\right) \cdot \frac{\partial v_t}{\partial \rho} = \\ &= \frac{\overline{v} - v_t}{\overline{v}} \left(1 - \frac{\rho}{\overline{v}} \frac{y}{1 + y\frac{\rho}{\overline{v}}}\right) > 0\end{aligned}$$

Subscription: Proving strict monotonicity for the subscription case is equivalent to the two-part-tariff case, with y given by:

$$y = \frac{\alpha \overline{v}}{\left(1 - \rho\left(\frac{\overline{v} - v_s}{\overline{v}}\right)\right)^2}$$

Pay-per-use: Proving strict monotonicity for the pay-per-use case is equivalent to the two-part-tariff case, with y given by:

$$y = \frac{\alpha \overline{v}}{\left(1 - \rho\left(\frac{\overline{v} - v_p}{\overline{v}}\right)\right)^3}$$

and the result follows. ■

2 Proofs from the Industry Standard System Time Model

With a two-part tariff the firm can adjust its price so that the consumer with value v is the indifferent consumer, yielding an arrival rate of $\Lambda \overline{F}\left(v\right)$ to the firm. Capacity still adjusts to meet the standard, so $\mu(v) = I + \Lambda \overline{F}\left(v\right)$. The expected time in system is I, so the consumer incurs congestion costs at rate w/I. The firm sets the subscription rate so that all consumers are indifferent between purchasing or not. Therefore, the firm's profit function is

$$\Pi_t(v) = \Lambda \overline{F}\left(v\right) \left(E\left[V | V \geq v\right] - \frac{w}{I}\right) - c\mu(v) :$$

the first term is the consumer's surplus when v is the indifferent consumer and the second term is the capacity cost. The following proposition establishes that an optimal threshold, v_t, exists and is unique.

Proposition 5 *The two-part tariff function $\Pi_t(v)$ is quasi-concave and $v_t = \arg\max_v \Pi_t(v)$ is uniquely defined by*

$$v_t = \frac{w}{I} + c \tag{8}$$

Proof. That (8) is a local maximum is shown by examining the first-order conditions of $\Pi_t(v)$.

$$\frac{d\Pi_t(v)}{dv} = -\Lambda f(v) \left(v - \frac{w}{I} - c\right)$$

Let $\varrho(v) = d\Pi_t(v)/dv$. Noting that $\varrho(0) > 0$ and that $\lim_{v\to\infty} \varrho(v) \leq 0$, indicates that there exists at least one maximum. Uniqueness is guaranteed because the RHS of (8) is constant in v. ■

Proposition 6 *$\Pi_x(v_x)$, $x \in \{s, p\}$ is convex-concave in I if $h'(v)/(h(v))^2$ is decreasing in v, where $h(v)$ is the hazard rate, $h(v) = f(v)/\bar{F}(v)$.*

Proof. Per-use: We first find a sufficient condition under which $\partial v_p/\partial I$ is increasing in I. Using the implicit function theorem, we get:

$$\frac{\partial v_p}{\partial I} = -\frac{w/I^2}{1+\frac{h'(v)}{(h(v))^2}} < 0. \tag{9}$$

Thus, we get that $\partial v_p/\partial I$ is increasing in I, if $h'(v)/(h(v))^2$ is decreasing in v (and hence, increasing in I). Applying the envelope theorem and differentiating $\Pi_p(v_p;I)$ with respect to I, we get:

$$\frac{\partial \Pi_p(v_p;I)}{\partial I} = \frac{\Lambda w \bar{F}(v_p)}{I^2} - c \tag{10}$$

Let $\varrho(I) = \partial \Pi_p(v_p;I)/\partial I$. Note that $\lim_{v\to 0}\varrho(v) > 0$ and that $\lim_{v\to\infty}\varrho(v) = -c$. To prove that $\Pi_p(v_p;I)$ is convex-concave in I is equivalent to showing that there exists a unique I that solves $\Pi_p''(v_p;I) = 0$. Taking the second derivative, equating to 0 and rearranging, we get:

$$I\frac{\partial v_p}{\partial I} = -\frac{2\bar{F}(v_p)}{f(v_p)}.$$

Given the sufficient condition above, because F is IFR, the term on the RHS is decreasing in I and the term on the LHS is increasing in I, providing the desired result.

Subscription: Following the same steps as in the per-use case, we find:

$$\frac{\partial v_s}{\partial I} = -\frac{w}{I^2} < 0 \tag{11}$$

which is increasing in I. Differentiating $\Pi_s(v_s;I)$ with respect to I, we get:

$$\frac{\partial \Pi_s(v_s;I)}{\partial I} = \frac{\Lambda w \bar{F}(v_s)}{I^2}(1-ch(v_s)) - c \tag{12}$$

Let $\varrho(I) = \partial \Pi_s(v_s;I)/\partial I$. Note that $\lim_{v\to 0}\varrho(v) > 0$ and that $\lim_{v\to\infty}\varrho(v) = -c$. To prove that $\Pi_s(v_s;I)$ is convex-concave in I is equivalent to showing that there exists a unique I that solves $\Pi_s''(v_s;I) = 0$. Taking the second derivative, equating to 0 and rearranging, we get:

$$I\frac{\partial v_s}{\partial I}\left(\frac{h'(v_s)}{h^2(v_s)} + \frac{1}{ch(v_s)} - 1\right) = -\frac{2(1-ch(v_s))}{ch^2(v_s)}.$$

Because $\partial v_s/\partial I$ is increasing in I, F is IFR and given the sufficient condition above, all terms in the LHS are increasing in I while the term in the RHS is decreasing in I, so the function is convex-concave in I. ■

Proposition 7 *If* $\overline{F}(v_p) < \overline{F}(v_s)(1-ch(v_s))$, *then* $\partial\Pi_p/\partial I < \partial\Pi_s/\partial I$. *Furthermore, if condition (A1) holds, then there exists a unique* $\widetilde{I}$ *such that* $\partial\Pi_p/\partial\widetilde{I} = \partial\Pi_s/\partial\widetilde{I}$.

Proof. Fix I. Comparing equations (10) and (12) we get the first result. To show uniqueness, consider $\partial\Pi_p/\partial\widetilde{I} = \partial\Pi_s/\partial\widetilde{I}$. Rearranging the condition, we have:

$$1 - ch(v_s) = \frac{\overline{F}(v_p)}{\overline{F}(v_s)}.$$

Because F is IFR, the LHS is increasing in I. It remains to show that $\overline{F}(v_p)/\overline{F}(v_s)$ is decreasing in I. Differentiating $\overline{F}(v_p)/\overline{F}(v_s)$ with respect to I, we get:

$$\frac{\partial}{\partial I}\left(\frac{\overline{F}(v_p)}{\overline{F}(v_s)}\right) = -\frac{\overline{F}(v_p)}{\overline{F}(v_s)}\left[h(v_p)\frac{\partial v_p}{\partial I} - h(v_s)\frac{\partial v_s}{\partial I}\right].$$

To show that $\overline{F}(v_p)/\overline{F}(v_s)$ is decreasing in I is equivalent to showing that the bracketed term is positive. Because $v_s < v_p$ and F is IFR, $h(v_p) > h(v_s)$. Thus, it is enough to show that $\partial v_p/\partial I > \partial v_s/\partial I$. From (9) and (11), we have:

$$\frac{\partial v_p}{\partial I} - \frac{\partial v_s}{\partial I} = -\frac{w}{I^2}\left(\frac{1}{1+\frac{h'(v)}{(h(v))^2}} - 1\right)$$

which is positive if (A1) holds. ■

Proposition 8 *When $V \sim U[0,\overline{v}]$, $\Pi_s > \Pi_p$ if $w/I \leq \overline{v} - \left(\sqrt{2}+1\right)c$.*

Proof. The subscription profit function is given by:

$$\Pi_s = \Lambda\overline{F}(v_s)\left(E[V|V \geq v_s] - v_s - c\right) - cI$$

The pay-per-use profit function is given by

$$\begin{aligned}\Pi_p &= \Lambda\bar{F}(v_p)\left(v_p - \frac{w}{I} - c\right) - cI = \\ &= \Lambda\bar{F}(v_p)\left(v_p - v_s - c\right) - cI\end{aligned}$$

where the second equality holds because $v_s = w/I$. $\Pi_s > \Pi_p$ when the following condition holds:

$$\overline{F}(v_s)\left(E[V|V \geq v_s] - v_s - c\right) > \bar{F}(v_p)\left(v_p - v_s - c\right)$$

Substituting $v_p = \frac{\overline{v}+v_s+c}{2}$ for the uniform and applying some algebra gives the desired result. ■

3 Proofs from the Endogenous Capacity Model

Like with a fixed capacity, with a two-part tariff the firm's revenue is $R_t(v)$, where v is the value of the consumer who, at a service opportunity, is indifferent between seeking service or not. By adding in the cost of capacity, the firm's profit function is

$$\Pi_t(v,\mu) = \Lambda\bar{F}(v)\left(E[V|V \geq v] - \frac{w}{\mu - \Lambda\bar{F}(v)}\right) - c\mu.$$

As with per-use pricing, $\Pi_t(v,\mu)$ is concave in μ for a fixed v and $\mu_t(v) = \mu_p(v)$ is the optimal capacity. The profit function can then be written as

$$\begin{aligned}\Pi_t(v) &= \Pi_t(v,\mu_t(v)) \\ &= \Lambda\bar{F}(v)\left(E[V|V \geq v] - c - \frac{2\phi}{\sqrt{\bar{F}(v)}}\right)\end{aligned}$$

The next proposition characterizes the firm's optimal policy. The proposition holds for any increasing generalized failure rate (IGFR) distribution[1].

Proposition 9 *If $\bar{v} > c$, there exists an upper bound $\overline{\phi}$, such that for every $\phi \leq \overline{\phi}$ there exists a unique interior optimal threshold level, $v_t = \arg\max_v \Pi_t(v)$, which is implicitly defined by the smaller of two possible solutions to:*

$$v_t = \frac{\phi}{\sqrt{\bar{F}(v_t)}} + c. \tag{13}$$

The optimal capacity is

$$\mu_t = \frac{wv_t}{(v_t - c)^2}, \tag{14}$$

and the optimal pricing parameters are

$$\begin{aligned} p_t &= c \\ k_t &= \bar{F}(v_t)\left(E\left[v|v \geq v_t\right] - v_t\right) \end{aligned}$$

Proof. Let

$$z(v) = \frac{\phi}{\sqrt{\bar{F}(v)}} + c.$$

Differentiating, we get:

$$z'(v) = \frac{\phi}{2\sqrt{\bar{F}(v)}} \cdot \frac{f(v)}{\bar{F}(v)}.$$

Notice that $z'(v) > 0$. Because $v_t = z(v_t)$, we have:

$$\frac{\phi}{\sqrt{\bar{F}(v_t)}} = v_t - c.$$

$z'(v_t)$ thus becomes:

$$z'(v_t) = \frac{v_t - c}{2} \cdot \frac{f(v_t)}{\bar{F}(v_t)}$$

Next, we show that $z'(v_t)$ is increasing by showing that $z''(v_t) > 0$.

$$z''(v_t) = \frac{1}{2} \frac{\left[\left(f(v_t) + (v_t - c) f'(v_t)\right) \bar{F}(v_t) + (v_t - c) f^2(v_t)\right]}{\left(\bar{F}(v_t)\right)^2}$$

$z''(v_t) > 0$ iff the bracketed term $\left(f(v_t) + (v_t - c) f'(v_t)\right) \bar{F}(v_t) + (v_t - c) f^2(v_t) > 0$. Rearranging terms, the condition becomes:

$$\frac{v_t}{v_t - c} f(v_t) \bar{F}(v_t) + v_t f^2(v_t) > -v_t f'(v_t) \bar{F}(v_t) \tag{15}$$

Because F is IGFR, it follows that:

$$f(v) \bar{F}(v) + v f^2(v) > -v f'(v) \bar{F}(v).$$

[1] F exhibits an increasing generalized failure rate (IGFR) iff $xh(x)$ is increasing, where $h(x) = f(x)/\bar{F}(x)$ is the failure rate. That is, the IGFR property is more general than the IFR– distributions that exhibit IFR are also IGFR, but not vice versa.

This property of IGFR distributions and the fact that we must have $v_t > c$, ensures that condition (15) holds. Therefore, $z''(v_t) > 0$. Given that $z(v)$ is increasing and $z'(v_t) = 1$ for at most one v_t, it follows that there can be only one v_t that solves $v_t = z(v_t)$.

Given that $\Pi(0)$ is finite and $\lim_{v\to\bar{v}} \Pi(v) = 0$, a maximum exists if there exists an $v_t < \bar{v}$ such that $\Pi'(v_t) = 0$ and $\Pi(v_t) \geq 0$. Requiring that $\Pi(v_t) \geq 0$ is equivalent to having

$$\frac{\Pi(v)}{\Lambda \bar{F}(v)} = E[V|V \geq v] - c - 2\phi\sqrt{\frac{1}{\bar{F}(v)}} \geq 0$$

for some v. Assume $\phi = 0$. If $\bar{v} > c$, there is a solution with positive profit. Let $M(\phi) \equiv \Pi(v_t(\phi), \phi)$. From the Envelope Theorem, we have:

$$\frac{\partial M(\phi)}{\partial \phi} = -2\Lambda\sqrt{\bar{F}(v_t)} < 0$$

which means that $\Pi(v_t(\phi), \phi)$ is decreasing in ϕ. This implies there exists some $\bar{\phi}$ such that $\Pi(v_t(\phi), \phi) \geq 0$ for $\phi \leq \bar{\phi}$. Otherwise, there does not exist an optimal $v_t < \bar{v}$.

Optimal pricing parameters: since customers are individual maximizers, customers care only about their own disutility of waiting while choosing whether or not to request service. For any given service rate level μ, customers' indifference is determined by a threshold valuation level v_e which satisfies: $v_e = wW\left(\Lambda \bar{F}(v_e)\right)$. If the monopolist chooses a service rate μ_t, the desired social optimal valuation threshold v_t is given by equation (13) and is equivalent to $v_e + c$. Thus choosing $p_t = c$ as the per use fee guarantees the social optimal congestion. Customer's expected utility under p_t equals $k_t = \int_{v_t}^{\infty} v dF(v) - \bar{F}(v_t)(v_e + c)$. Setting the subscription rate to k_t enables the monopolist to extract the remaining welfare and together with the per-use price p_t and the service rate cost $c\mu_t$, his expected profit rate results in the social optimal welfare $\Pi(v_t)$. ■

As with fixed capacity, the firm is able to extract all consumer welfare, so the two-part tariff is optimal for the firm (again, assuming only a single price can be charged per transaction).

Proposition 10 *Compared to the social optimal, the following hold:*

1. *In pay-per-use, the system is under congested and there is lower investment in capacity.*
2. *If $V \sim U[0, \overline{v}]$ subscription results in over congestion ($v_s = v_t - c$) and in higher investment in capacity.*

Proof. 1. Pay-per-use. Congestion: compare the thresholds given in (13) and

$$v_p = \frac{\bar{F}(v_p)}{f(v_p)} + \frac{\phi}{\sqrt{\bar{F}(v_p)}} + c. \tag{16}$$

Let $g_t(v) = \sqrt{cw/\Lambda\bar{F}(v)} + c$ and $g_p(v) = \bar{F}(v)/f(v) + \sqrt{cw/\Lambda\bar{F}(v)} + c$. Since $\bar{F}(v)/f(v) > 0$, $g_p(v) > g_t(v)$, $\forall v$. Thus, it must be that $v_t < v_p$. Capacity: We have previously established that the optimal service rate function $\mu(v)$ in

$$\mu(v) = \Lambda\bar{F}(v) + \sqrt{\frac{w\Lambda\bar{F}(v)}{c}}$$

is the same for both pay-per-use and the social welfare. Since $\mu'(v) < 0$, it follows that $\mu(v_t) > \mu(v_p)$.

2. Subscription. Congestion: Assume $v_t = v_s + c$ holds and check that the first-order conditions given in (13) are satisfied. For the uniform case, condition (13) becomes:

$$v_t = \phi\sqrt{\frac{\overline{v}}{(\overline{v} - v_t)}} + c$$

Substituting $v_t = v_s + c$, we get

$$v_s = \phi\sqrt{\frac{\overline{v}}{(\overline{v} - v_s - c)}}$$

which is exactly condition

$$v_s = \phi\sqrt{\frac{1}{\bar{F}(v_s) - cf(v_s)}}. \tag{17}$$

applied for the uniform distribution. Capacity: For the uniform case, we get:

$$\mu(v_s) = \frac{w}{v_s} + \frac{\Lambda(\overline{v} - v_s)}{\overline{v}}$$

and

$$\mu(v_t) = \frac{wv_t}{(v_t - c)^2} = \frac{w}{v_s} + \frac{wc}{v_s^2}$$

where the last equality follows from substituting $v_t = v_s + c$. Now, $\mu(v_s) > \mu(v_t)$ if and only if

$$\phi^2 < \frac{v_s^2(\overline{v} - v_s)}{\overline{v}}. \tag{18}$$

The condition must hold to get an interior solution to the problem. To see this, note that $\bar{\phi} = (v_t^o - c)\sqrt{\bar{F}(v)}$ for $V \sim U[0, \overline{v}]$, where v_t^o is the social optimal threshold level that solves both $\Pi(v) = 0$ and $\Pi'(v) = 0$. Because we must have $\phi < \bar{\phi}$ for an interior solution, it follows that

$$\phi < (v_t^o - c)\sqrt{\frac{\overline{v} - v_t^o}{\overline{v}}}$$

As $v_t = v_s + c$, we get:

$$\phi^2 < \frac{v_s^2(\overline{v} - v_t^o)}{\overline{v}}.$$

and because $v_s < v_t^o$, condition (18) follows. ■

Proposition 11 *The following inequalities hold: (i) $u_t(c) > u_p(c)$ $\forall c$ (ii) if $V \sim U[0, \overline{v}]$, $u_s(c) > u_t(c)$ $\forall c$.*

Proof. Let $u_x(c) = \Lambda\bar{F}(v_x)/\mu_x$, $x \in \{s, t, p\}$. *(i)* Theorem 6 and Proposition 9 imply:

$$\frac{1}{u_x} = 1 + \sqrt{\frac{w}{c\Lambda\bar{F}(v_x)}} \quad x \in \{t, p\}.$$

The result follows for every IFR distribution, because $v_t < v_p$ (Proposition 10).

(ii) Let $V \sim U[0, \overline{v}]$. We have:

$$\frac{1}{u_s} = \frac{w\overline{v}}{\Lambda v_s(\overline{v} - v_s)} + 1 \tag{19}$$

and

$$\begin{aligned}\frac{1}{u_t} &= \frac{wv_t}{\Lambda\left(v_t-c\right)^2}\cdot\frac{\overline{v}}{\overline{v}-v_t}\\ &= \frac{w\overline{v}\left(v_s+c\right)}{\Lambda v_s^2\left(\overline{v}-v_s-c\right)}\\ &= \frac{w\overline{v}}{\Lambda v_s\left(\overline{v}-v_s-c\right)}+1 \qquad (20)\end{aligned}$$

where the first equality follows from equation (14), the second equality follows because $v_s = v_t - c$ (Proposition 10) and the third equality from substituting in (17). Comparing (19) and (20), the result follows. ■

Proposition 12 *Let $\widetilde{w}$ be the maximum w such that $\Pi_p(w) \geq 0$ and $\Pi_p(w) \leq \Pi_s(w)$. If $\overline{v} > \left(1+\sqrt{2}\right)c$, then either there exists a unique threshold $\widetilde{w}$ such that for $w < \widetilde{w}$ it is better to sell subscriptions and for $w > \widetilde{w}$, using pay-per-use results in higher profits, or subscription is always better than pay-per-use, conditional on obtaining positive profits. Otherwise, if $c < \overline{v} < \left(1+\sqrt{2}\right)c$, pricing on the basis of pay-per-use is always better than selling subscriptions.*

Proof. Observe first, that at $w = 0$, we have $v_s = 0$ and $v_p = \bar{F}(v_p)/f(v_p) + c = \left(\overline{v}+c\right)/2$ (when the second inequality follows from substituting for the uniform distribution). These lead to $\Pi_s\left(v_s; w=0\right) = \Lambda\left(\int_0^\infty v dF(v) - c\right) = \Lambda\left(E[V]-c\right)$ and to $\Pi_p\left(v_p; w=0\right) = \Lambda\bar{F}(v_p)\left(v_p-c\right) = \left(\overline{v}-c\right)^2/4\overline{v}$, respectively. Comparing between the profit functions, we get that $\Pi_s\left(v_s; w=0\right) \geq \Pi_p\left(v_p; w=0\right)$ iff $\overline{v} > \left(1+\sqrt{2}\right)c$. Rewriting equation (16) for the uniform distribution, we get:

$$\left(2v_p-\overline{v}\right)\left(1-\frac{v_p}{\overline{v}}\right) = c+\sqrt{cw}.$$

Similarly, from equation (13) we get:

$$v_s\sqrt{1-\frac{v_s}{\overline{v}}-\frac{c}{\overline{v}}} = \sqrt{cw}.$$

Let $M_s(w) \equiv \Pi_s\left(v_s; w\right)/\Lambda$ and $M_p(w) \equiv \Pi_p\left(v_p; w\right)/\Lambda$. Substituting v_p and v_s, we have:

$$M_p(w) = \left(1-\frac{v_p}{\overline{v}}\right)\left(v_p-c\right) - 2\sqrt{cw\left(1-\frac{v_p}{\overline{v}}\right)}$$

and

$$M_s(w) = \frac{\overline{v}}{2}\left(1-\frac{v_s}{\overline{v}}\right)^2 - \sqrt{cw\left(1-\frac{v_s}{\overline{v}}-\frac{c}{\overline{v}}\right)} - c\left(1-\frac{v_s}{\overline{v}}\right)$$

By the Envelope Theorem, we first find that both profit functions are decreasing in w:

$$M_p'(w) = -w^{-\frac{1}{2}}\sqrt{c\left(1-\frac{v_p}{\overline{v}}\right)}$$

and

$$M_s'(w) = -\frac{1}{2}w^{-\frac{1}{2}}\sqrt{c\left(1-\frac{v_s}{\overline{v}}-\frac{c}{\overline{v}}\right)} = -\frac{1}{2}w^{-\frac{1}{2}}\sqrt{c\left(1-\frac{v_t}{\overline{v}}\right)}$$

where the second equality follows from Proposition 10. Thus, both profit functions are decreasing in w. Applying some algebra along with the fact that $v_t < v_p$, we get that $M_s'(w) < M_p'(w) < 0$

$\forall w$. Combining this with the condition at $w = 0$, the three cases follow. ■

Proposition 13 *Let $\widetilde{c}$ be the maximum c such that $\Pi_p(c) \geq 0$ and $\Pi_p(c) \leq \Pi_s(c)$. Then either there exists a unique capacity cost threshold $\widetilde{c}$, below which it is better to use subscription and above which using per use pricing results in higher profits or subscription is always better than pay-per-use, conditional on obtaining positive profits*[2]*.*

Proof. Observe first, that at $c = 0$, we have $v_s = 0$ and $v_p = \bar{F}(v_p)/f(v_p)$. These lead to $\Pi_s(v_s; c = 0) = \Lambda E[V]$ and to $\Pi_p(v_p; c = 0) = \Lambda \bar{F}(v_p) v_p$, respectively. Because $\Pi_s(v_s; c = 0) = \Pi_t(v_t; c = 0)$, we must have $\Pi_p(v_p; c = 0) \leq \Pi_s(v_s; c = 0)$. Let $M_s(c) \equiv \Pi_s(v_s; c)$ and $M_p(c) \equiv \Pi_p(v_p; c)$. By the Envelope Theorem, $M_s'(c) = -\mu_s < 0$ and $M_p'(c) = -\mu_p < 0$. Using the assumption that $\mu_s > \mu_p$, we get that $|M_s'(c)| > \left|M_p'(c)\right| \ \forall c$. From this and the fact that $\Pi_s(v_s; c = 0) > \Pi_p(v_p; c = 0)$, the result follows. Because it might be that $\Pi_s(v_s; \widetilde{c}) = \Pi_p(v_p; \widetilde{c}) < 0$, and we are only interested in the non-negative range of the profit functions, subscription might be better than pay-per-use in the entire relevant range. ■

4 Extension to Linear Waiting Costs

In this section we assume that customers' waiting costs are a function of the value of the service. More specifically, we assume that waiting costs are linear in the value, i.e. $w(v) = a + bv$. In the base models we assumed that $b = 0$, but consumers' patience may depend on the value they attach to the service. Customers may be more patient ($b < 0$) or less patient ($b > 0$) as their value increases. In what follows, we analyze the exogenous capacity case under this assumption. As noted in the main text, the results obtained under constant waiting costs generalize when we allow for $b \neq 0$.

4.1 Per-Use Pricing

Given that p, a, b and λ are common to all customers (they all have the same expectations) and constant across time, there is some threshold value, v, such that a customer seeks service whenever the realized value of an opportunity is v or greater, and otherwise the customer passes on the opportunity:

$$v = p + (a + bv) W(\lambda).$$

The actual arrival rate to the service is then $\Lambda \bar{F}(v)$. The next proposition establishes the existence and uniqueness of the equilibrium per-use threshold.

Proposition 14 *The per-use threshold value, v, given by:*

$$v = p + (a + bv) W(\Lambda \bar{F}(v)) \tag{21}$$

exists and is unique.

Proof. For expectations to be consistent with actual operating conditions (i.e., $\lambda = \Lambda \bar{F}(v)$) the threshold v must satisfy (21). To show uniqueness, consider two cases. If $b < 0$, uniqueness is

[2] This result holds in general for all IFR distributions assuming that $\mu_s > \mu_p$ for all c. Note that the condition $\mu_s > \mu_p$ holds for the uniform distribution (Proposition 10) and is shown to hold numerically for the Weibull distribution (with a wide range of parameter combinations) as well (see the numerical study in §6 of this appendix).

guaranteed, because the LHS is increasing in v and the RHS is decreasing. If $b > 0$, rearrange (21) as:

$$v(1 - bW(\Lambda\bar{F}(v))) = p + aW(\Lambda\bar{F}(v)).$$

Because $b > 0$, the LHS is increasing in v and the RHS is decreasing in v, so the threshold is unique. To show existence, note that v must be positive. This implies that we must have $1 > bW(\Lambda\bar{F}(v))$ or $b < 1/W(\Lambda\bar{F}(v))$. The condition holds for all v, because we must have $b < 1/W(\Lambda)$ (otherwise no customer joins) and W is increasing. ■

The firm's revenue is $R_p = \lambda p$, which can be expressed in terms of the threshold v:

$$R_p(v) = \Lambda F(v)(v - (a + bv)\, W(\Lambda F(v))).$$

The following proposition establishes that an optimal threshold, v_p, exists and is unique.

Proposition 15 *The per-use revenue function, $R_p(v)$, is quasi-concave and $v_p = \arg\max_v R_p(v)$ is uniquely defined by*

$$v_p = \frac{\bar{F}(v_p)}{f(v_p)}\left(1 - bW(\Lambda\bar{F}(v_p))\right) + (a + bv)\, W(\Lambda\bar{F}(v_p)) + (a + bv)\,\Lambda\bar{F}(v_p)W'(\Lambda\bar{F}(v_p)). \tag{22}$$

Proof. That (22) is a local maximum, is shown by examining the first-order conditions of $R_p(v)$.

$$\begin{aligned}\frac{dR_p(v)}{dv} &= -\Lambda f(v)\left(v - (a + bv)\, W(\Lambda\bar{F}(v))\right) + \\ &\quad +\Lambda\bar{F}(v)\left(1 + (a + bv)\,\Lambda f(v)W'(\Lambda\bar{F}(v))\right) - bW(\Lambda\bar{F}(v)))\end{aligned}$$

Let $\varrho(v) = dR_p(v)/dv$. Noting that $\varrho(0) > \Lambda$ (because $1 > bW(\Lambda)$) and that $\lim_{v\to\infty}\varrho(v) \leq 0$, indicates that there exists at least one maximum. Showing that the solution v_p is unique will complete the proof. v_p satisfies the first-order condition given by (22). Consider $b < 0$ and $b > 0$ separately. If $b < 0$, the LHS is increasing and the RHS is decreasing ensuring a unique v_p. If $b > 0$, rearranging the terms in (22) gives:

$$\begin{aligned}&v_p\left(1 - b\Lambda\bar{F}(v_p)W'(\Lambda\bar{F}(v_p))\right) - bW(\Lambda\bar{F}(v_p))v_p\left(1 - \frac{\bar{F}(v_p)}{v_p f(v_p)}\right) \\ = \;& aW(\Lambda\bar{F}(v_p)) + \frac{\bar{F}(v_p)}{f(v_p)} + a\Lambda\bar{F}(v_p)\, W'\left(\Lambda\bar{F}(v_p)\right)\end{aligned}$$

Since F is IFR, the LHS is increasing in v_p and the RHS is decreasing in v_p, guaranteeing uniqueness. ■

4.2 Subscription Pricing

Under subscription, in equilibrium, the indifferent consumer's value, v_s, exactly equals the expected congestion cost:

$$v_s = (a + bv_s)\, W\left(\Lambda\bar{F}(v_s)\right). \tag{23}$$

Proving the existence and uniqueness of the subscription threshold value follows the same lines of Proposition 14 and will not be repeated.

4.3 Comparison

This section compares the revenues generated by the two pricing schemes. To make these comparisons more explicit, we assume in this section $V \sim U[0, \overline{v}]$ and $W(\lambda) = 1/(\mu - \lambda)$. The following proposition generalize the comparative results from the fixed capacity case to allow for linear waiting costs.

The next proposition defines the set of parameters for which the firm can earn non-negative revenue. Although the firm's problem is determined by five parameters (a, b, μ, Λ and $\overline{v}$), the next theorem indicates that the pricing schemes' relative rankings depend only on three of them.

Proposition 16 *[generalization of Lemma 1] The optimal revenue with each pricing scheme (R_p and R_s) can be expressed in terms of α, β, ρ, Λ and $\overline{v}$, where $\alpha = a/\mu\overline{v}$, $\beta = b/\mu$ and $\rho = \Lambda/\mu$. Furthermore, the two revenues are non-negative for $(\alpha + \beta) \in [0, 1]$ and linear in $\Lambda\overline{v}$. Consequently, the relative revenue, R_s/R_p, can be expressed in terms of α, β and ρ.*

Proof. For the per-use case, we have established in Proposition 15 that $R_p(v)$ is quasi-concave. Also note that $R_p(\overline{v}) = 0$. $\overline{v}$ will be the maximizer, when the first-order conditions at $\overline{v}$ are non negative. Evaluating the first-order conditions at $v_p = \overline{v}$ we get: $-\Lambda\left(\overline{v} - (a + b\overline{v})/\mu\right)/\overline{v} \geq 0$, which is equivalent to $\alpha + \beta \geq 1$. Since $R_p(\overline{v}) = 0$, this will imply that $R_p(v) < 0 \; \forall v \in [0, \overline{v})$. Thus, we can limit our search space to the interesting case, $\alpha + \beta < 1$. In this case there exists an interior solution that results in a positive per-use revenue rate. Performing a similar analysis for the subscription case, reveals that the condition for an interior maximum that guarantees positive revenue rates is the same as in the per-use case.

Next, we show that for any given α, β and ρ, the revenue functions are linear in $\overline{v}$, which implies that $\overline{v}$ does not affect the comparison between subscription and pay per use. The comparison can be made solely on the basis of α, β and ρ. To see that this is indeed the case, note that $v_s/\overline{v}$ can be implicitly expressed by:

$$\frac{v_s}{\overline{v}} = \frac{\alpha + \beta\frac{v_s}{\overline{v}}}{1 - \rho\left(1 - \frac{v_s}{\overline{v}}\right)} \tag{24}$$

Thus, $v_s = \overline{v} \cdot g(\alpha, \beta, \rho)$, where g is a function of α, β and ρ only. Plugging v_s into the subscription revenue function, we obtain:

$$R_s = \frac{\Lambda\left(1 - g(\alpha, \beta, \rho)\right)^2}{2} \cdot \overline{v}$$

which is linear in $\overline{v}$. Similarly for the pay-per-use case, we can express $v_p/\overline{v}$ by:

$$\frac{v_p}{\overline{v}} = \left(1 - \frac{v_p}{\overline{v}}\right)\left(1 + \frac{(\alpha + \beta)\rho - \beta}{\left(1 - \rho\left(1 - \frac{v_p}{\overline{v}}\right)\right)^2}\right) + \frac{\alpha + \beta\frac{v_p}{\overline{v}}}{1 - \rho\left(1 - \frac{v_p}{\overline{v}}\right)} \tag{25}$$

Thus, $v_p = \overline{v} \cdot h(\alpha, \rho)$, where h is a function of α, β and ρ only. Plugging v_p into the subscription revenue function, we obtain:

$$R_p = \Lambda\left(1 - h(\alpha, \beta, \rho)\right)\left(h(\alpha, \beta, \rho) - \frac{\alpha + \beta h(\alpha, \beta, \rho)}{1 - \rho\left(1 - h(\alpha, \beta, \rho)\right)}\right) \cdot \overline{v}$$

which is also linear in $\overline{v}$. Thus, to compare the revenues of the two schemes it suffices to examine changes in α, β and ρ. ■

Next, we show that there is a single crossing point for the two revenue functions, R_s and R_p. Proposition 17 establishes the result analytically for $b < 0$.

Proposition 17 *[generalization of Theorem 2] If $b < 0$, for each pair of values (α, β), there exists a unique $\widetilde{\rho}(\alpha, \beta)$, such that subscription yields higher revenue than per-use pricing for $\rho < \widetilde{\rho}(\alpha, \beta)$. Otherwise, per-use pricing yields higher revenue.*

Proof. Uniqueness: note first, that $\lim_{\rho \to 0} R_s(\rho) = \frac{\Lambda \overline{v}}{2} \left(\frac{1-\alpha-\beta}{1-\beta} \right)^2$ and $\lim_{\rho \to 0} R_p(\rho) = \frac{\Lambda \overline{v}(1-\alpha-\beta)^2}{4(1-\beta)}$. Also, $\lim_{\rho \to \infty} R_s(\rho) = \lim_{\rho \to \infty} R_p(\rho) = 0$. Finally, we show that both R_s and R_p are monotonically decreasing in ρ and that $R_s'(\rho) < R_p'(\rho)\ \forall \rho$. By implicitly differentiating the revenue functions, we get:

$$\frac{dR_s(\rho)}{d\rho} = -\frac{(1-g)^2(\alpha + \beta g)}{(1-\rho(1-g))^2}$$

and

$$\frac{dR_p(\rho)}{d\rho} = -\frac{(1-h)^2(\alpha + \beta h)}{(1-\rho(1-h))^2}$$

where g and h are shorthand notation for $g(\alpha, \beta, \rho)$ and $h(\alpha, \beta, \rho)$. To establish that $R_s'(\rho) < R_p'(\rho)\ \forall \rho$, note that $v_s(\rho) < v_p(\rho)\ \forall \rho$. The result follows immediately if $b < 0$ (note, $\beta = b/\mu$, so $b < 0 \Leftrightarrow \beta < 0$). To see that, let

$$z(x) = -\frac{(1-x)^2(\alpha + \beta x)}{(1-\rho(1-x))^2}.$$

Observe that if $\beta < 0$, both terms in the numerator are decreasing in x and the term in the denominator is increasing in x, so that $z'(x) > 0$. Thus, because $g < h$, we must have that $R_s'(\rho) < R_p'(\rho)\ \forall \rho$. ■

While we were unable to show that there is a single crossing point when $b > 0$, the result does hold numerically. Because we must have $\alpha + \beta < 1$ to achieve positive profits, we were able to cover the entire parameter space by choosing different values of $\alpha > 0$, $\beta > 0$ such that the condition holds. Table 1 provides the potential utilization rate, $\widetilde{\rho}(\alpha, \beta)$, at which the two schemes yield the same revenue. For each pair of values (α, β) in the table (and for 60 additional (α, β) pairs that satisfy $\alpha + \beta < 1$ and $\beta \geq 0$), we demonstrate numerically that there exists a unique $\widetilde{\rho}(\alpha, \beta)$ that yields the same revenue for both pricing schemes. This implies that, analogously to the fixed waiting cost case, subscription pricing is the preferable pricing scheme for all $\rho < \widetilde{\rho}(\alpha, \beta)$ whereas per-use is preferable if $\rho > \widetilde{\rho}(\alpha, \beta)$.

5 Analysis of Mixed Strategy Equilibria

In this section we extend the analysis of the three subscription models to allow for mixed strategy equilibria. Under pure strategies, the monopolist charges a subscription price k and makes all M customers buy. If we allow for mixed strategy equilibria, the monopolist can charge a subscription price $k(\gamma)$ such that a fraction γ of the customers purchase and $1 - \gamma$ do not ($\gamma \in [0, 1]$). In what follows we show that allowing for a mixed strategy equilibrium in which the firm charges $k(\gamma)$ and a fraction $\gamma < 1$ of customers purchase subscription can be sustained in equilibrium when capacity is fixed and the service rate is low compared to the potential arrival rate (so that congestion is an issue). In this case, the firm can extract higher revenues by charging a high subscription fee so that only a fraction of the consumers buy. We also show that in all other cases the sustainable equilibrium is the one in pure strategies. We assume throughout the section that $w(\lambda) = 1/(\mu - \lambda)$ and that $V \sim U[0, \overline{v}]$.

Table 1. Potential utilization rates, $\widetilde{\rho}(\alpha,\beta)$, that yield identical revenue with per-use and subscription pricing, as well as actual utilizations when the potential arrival rate is $\widetilde{\rho}(\alpha,\beta)\mu$.

			Actual utilization (%) when $\rho=\widetilde{\rho}(\alpha,\beta)$	
α	β	$\widetilde{\rho}(\alpha,\beta)$	Per-use	Subscription
0.99	0.000	1.002	0.25	0.5
0.75	0.000	1.069	7.05	13.84
0.75	0.249	1.249	0.03	0.06
0.50	0.000	1.153	16.4	31.34
0.50	0.250	1.336	8.95	17.48
0.50	0.499	1.500	0.04	0.07
0.25	0.000	1.263	30.48	55.44
0.25	0.250	1.445	21.46	40.32
0.25	0.500	1.604	10.91	21.19
0.25	0.749	1.751	0.04	0.09
0.01	0.000	1.411	64.84	96.81
0.01	0.250	1.587	42.37	73.15
0.01	0.500	1.733	26.31	48.61
0.01	0.750	1.867	12.33	23.85
0.01	0.989	1.984	0.05	0.1

5.1 Fixed Capacity

Given that a fraction γ of the customers purchased a subscription, they will request service iff

$$v_s(\gamma) = wW\left(\gamma\Lambda\bar{F}(v_s(\gamma))\right) \tag{26}$$

Note that v_s is increasing in γ. As less customers purchase a subscription, the threshold value v_s according to which a subscribed customer requests service decreases as well and he requests service more often.

As part of the purchasing decision and given that a fraction γ of the customers purchase, a customer expects that a subscription generates the following net value per service opportunity,

$$\bar{F}(v_s(\gamma))\left(E\left[V|V\geq v_s(\gamma)\right]-v_s(\gamma)\right)$$

Given that service opportunities arrive at rate τ, the firm can choose which $k(\gamma)\geq k$ to set (where k is the subscription price charged in a pure strategy equilibrium, i.e. when $\gamma=1$). Each subscription price chosen corresponds to a unique fraction of customers, γ, buying a subscription.

$$k(\gamma) = \tau\bar{F}(v_s(\gamma))\left(E\left[V|V\geq v_s(\gamma)\right]-v_s(\gamma)\right)$$

The firm's resulting revenue can be expressed in terms of the fraction γ :

$$R_s(\gamma, v_s(\gamma)) = \gamma k(\gamma) M = \gamma\Lambda\bar{F}(v_s(\gamma))\left(E\left[V|V\geq v_s(\gamma)\right]-v_s(\gamma)\right) \tag{27}$$

where $v_s(\gamma)$ is given by (26). The firm can control the fraction of customers subscribing by changing

the subscription price, thereby also controlling congestion.

Proposition 18 *In the fixed capacity model there exists a unique equilibrium in which the firm charges a subscription price $k(\gamma)$ and a fraction γ of the customers purchase. Moreover, if $\Lambda \leq \mu$, then $\gamma = 1$ is always optimal–the resulting equilibrium is in pure strategies. However, if $\Lambda > \mu$ then the resulting equilibrium is in mixed strategies–customers purchase a subscription with probability $\gamma = 1/\rho$.*

Proof. Existence: total differentiating (27) with respect to γ, we get:

$$\begin{aligned}\frac{dR_s(\gamma, v_s(\gamma))}{d\gamma} &= \frac{\partial R_s(\gamma, v_s(\gamma))}{\partial \gamma} + \frac{\partial R_s(\gamma, v_s(\gamma))}{\partial v_s} \cdot \frac{dv_s}{d\gamma} \\ &= \Lambda \bar{F}(v_s(\gamma)) \left(E[V|V \geq v_s(\gamma)] - v_s(\gamma) - \gamma \frac{dv_s(\gamma)}{d\gamma} \right)\end{aligned}$$

where

$$\frac{dv_s(\gamma)}{d\gamma} = \frac{\Lambda \bar{F}(v_s(\gamma)) w W'\left(\gamma \Lambda \bar{F}(v_s(\gamma))\right)}{1 + \gamma \Lambda f(v_s(\gamma)) w W'\left(\gamma \Lambda \bar{F}(v_s(\gamma))\right)} > 0.$$

Let $\varrho(\gamma) = dR_s(\gamma)/d\gamma$. Noting that $\lim_{\gamma \to 0} \varrho(\gamma) > 0$ and that the domain of γ is closed from above, indicates that there exists at least one maximum.

Uniqueness: rewrite (26) in terms of γ:

$$\gamma(v_s) = \frac{\mu - \frac{w}{v_s}}{\Lambda \overline{F}(v_s)} \tag{28}$$

Plugging the expression in the revenue function, we get:

$$\begin{aligned}R_s(v_s) &= \gamma(v_s) k(v_s) M = \left(\mu - \frac{w}{v_s}\right)(E[V|V \geq v_s] - v_s) \\ &= \left(\mu - \frac{w}{v_s}\right)\left(\frac{\bar{v} - v_s}{2}\right)\end{aligned}$$

where the last inequality follows because of the uniform distribution assumption. Solving for the FOC yields:

$$v_s = \sqrt{\frac{w\bar{v}}{\mu}} \tag{29}$$

which is unique.

To find the conditions for which $\gamma < 1$ is optimal, plug equation (29) in (28). Simplifying, we obtain that $\gamma(v_s) = 1/\rho$. This γ is achievable if $\Lambda > \mu$. Otherwise, the optimal strategy is for the firm to charge a subscription price k such that all consumers purchase a subscription ($\gamma = 1$). ■

Note that the equilibrium found in the exogenous capacity model (§4 in the main text) represents a lower bound on firm's achievable revenues. If we allow for mixed strategy equilibria, given that $\Lambda > \mu$, the firm can do better by charging a higher subscription price and having a fraction $\gamma = 1/\rho$ of all customers purchase. Thus, it turns out that in the fixed capacity case subscription can do even better relative to per-use. If $\Lambda \leq \mu$ (i.e. if service rate is high relative to the potential arrival rate so that congestion is not problematic), the pure strategy equilibrium in which the firm charges a subscription fee k such that all customers purchase is, in fact, the optimal strategy for the firm.

5.2 Industry Standard System Time

The industry standard adjusts to the fraction of customers that join. In this case, $v_s(\gamma) = w/\left(\mu_s(\gamma) - \gamma\Lambda\overline{F}(v_s(\gamma))\right) = w/I$. The threshold is independent of γ because the firm adjusts the buffer capacity so that expected waiting time does not exceed the standard no matter what the fraction γ is. The firm's profit is then

$$\begin{aligned}\Pi_s(v_s(\gamma)) &= \gamma\Lambda\overline{F}(v_s)\left(E[V|V \geq v_s] - v_s\right) - c\mu(v_s) \\ &= \gamma\Lambda\overline{F}(v_s)\left(E[V|V \geq v_s] - v_s - c\right) - cI\end{aligned}$$

So the best the firm can do is to charge according to the pure strategy equilibrium ($\gamma = 1$).

5.3 Endogenous Capacity

Proposition 19 demonstrates that when capacity choice is endogenous, the sustainable equilibrium is one in pure strategies. That is, the optimal subscription price for the firm is the one found in §6 in the main text, which results in the entire customer base purchasing.

Proposition 19 *In the endogenous capacity case, if $E[V] > c$ and $\phi \leq \overline{\phi}_s$, then $\gamma = 1$ is always optimal–the resulting equilibrium is in pure strategies.*

Proof. Proposition 18 states that if $\mu < \Lambda$, then $\gamma(\mu) < 1$. In this case, the indifferent customer threshold satisfies $v_s(\mu) = \sqrt{w\bar{v}/\mu}$. However, if $\mu > \Lambda$, then $\gamma(\mu) = 1$, in which case v_s is implicitly defined by $v_s = w/\left(\mu - \Lambda\overline{F}(v_s)\right) = w/\left(\mu - \Lambda(\bar{v} - v_s)/\bar{v}\right)$. Given this optimal behavior for a fixed service rate, we are interested in finding the optimal service rate for the firm to invest in, μ_s. The firm's profit function is given by:

$$\begin{aligned}\Pi_s(\mu) &= \begin{cases} \gamma\Lambda\overline{F}(v_s)\left(E[V|V \geq v_s(\gamma)] - v_s(\gamma)\right) - c\mu(v_s(\gamma)) & \text{if } \mu < \Lambda \\ \Lambda\overline{F}(v_s)\left(E[V|V \geq v_s] - v_s\right) - c\mu & \text{if } \mu \geq \Lambda \end{cases} \\ &= \begin{cases} \frac{1}{2}\left(\mu - \sqrt{\frac{w\mu}{\bar{v}}}\right)\left(\bar{v} - \sqrt{\frac{w\bar{v}}{\mu}}\right) - c\mu & \text{if } \mu < \Lambda \\ \frac{\Lambda}{2\bar{v}}(\bar{v} - v_s)^2 - c\mu & \text{if } \mu \geq \Lambda \end{cases}\end{aligned}$$

where the last equality follows from the $M/M/1$ and uniform assumptions. It is easy to verify that Π_s is continuous and that it is strictly convex in the $\mu < \Lambda$ domain ($\partial\Pi_s^2(\mu)/\partial\mu^2 = \sqrt{w\bar{v}}/\left(4\mu^{3/2}\right)$). Moreover, from $v_s \leq \bar{v}$, we must have that $\mu \geq w/\bar{v}$ and $\Pi_s(w/\bar{v}) < 0$. In the domain $\mu \geq \Lambda$, the optimal service rate satisfies $\mu_s = \Lambda\overline{F}(v_s) + w/v_s$ (§6 in the main text). Provided that a positive profit can be made (i.e. if $E[V] > c$ and $\phi \leq \overline{\phi}_s$), $\mu_s \geq \Lambda$ if and only if $v_s \leq \sqrt{w\bar{v}/\Lambda}$. This condition always holds, because $v_s = w/\left(\mu - \Lambda\overline{F}(v_s)\right) = \sqrt{w\bar{v}/\Lambda}$ when $\mu = \Lambda$ and it is decreasing in μ. Taken together, continuity, $\Pi_s(\mu_s) > 0$ and the fact that $\mu_s \geq \Lambda$ imply that the optimal μ_s which maximizes the profit function under subscription involves the firm selling subscriptions to the entire customer base ($\gamma = 1$ is optimal) and it is chosen to be high enough so that $\mu_s > \Lambda$. In this proposition we assume that $E[V] > c$ and $\phi \leq \overline{\phi}_s$, which implies that the firm can make positive profits. If positive profit cannot be made, then the best the firm can do is to set $\mu_s = 0$, in which case nobody buys. ■

6 Comparison of Pricing Schemes when Capacity is Endogenous– A Numerical Study

Analytical comparison between v_t and v_s for distributions other than the uniform is difficult in the endogenous capacity case. While we were unable to prove that $v_s < v_t$ in general, an extensive numerical study confirmed our intuition, and demonstrated that it was the case for the Weibull distribution with $\kappa \geq 1$ as well. The following proposition restricts the number of parameters defining the problem.

Proposition 20 *Separate manipulation of w and Λ adds no additional value to the comparison of the pricing schemes over the manipulation of $\psi = w/\Lambda$*[3].

Proof. First note that comparing the profit rates $\Pi_p(v_p)$, $\Pi_s(v_s)$ and $\Pi_t(v_t)$ is equivalent to comparing $\Pi_p(v_p)/\Lambda$, $\Pi_s(v_s)/\Lambda$ and $\Pi_t(v_t)/\Lambda$, because we are only interested in the functions' relative ranking. Notice that the threshold expressions given in equations (13), (16) and (17) are all a function of ϕ. Moreover, the profit rate functions can be rewritten as:

$$\frac{\Pi_t(v_t)}{\Lambda} = \int_{v_t}^{\infty} v dF(v) - c\bar{F}(v_t) - 2\sqrt{c\psi\bar{F}(v_t)},$$

$$\frac{\Pi_p(v_p)}{\Lambda} = \bar{F}(v_p)\left(v_p - c - 2\sqrt{\frac{c\psi}{\bar{F}(v_p)}}\right),$$

and

$$\frac{\Pi_s(v_s)}{\Lambda} = \left(\int_{v_s}^{\infty} v dF(v) - (v_s + c)\bar{F}(v_s)\right) - \frac{c\psi}{v_s}$$

which are all functions of c and ψ and never of w or Λ alone. ■

Proposition 20 limited the number of parameters required for manipulation. We observed that it is not necessary to manipulate w and Λ separately, because they always appear as ratios both in the optimal thresholds and the profit functions. Only the ratio $\psi = w/\Lambda$ influences the ranking of the two pricing schemes.

We constructed 784 instances using all combinations of the parameters in Table 2 [4]. The values were chosen to cover a large parameter space. The Weibull distribution is completely characterized by the parameters κ and β, where the mean of the distribution is given by $\beta\Gamma(1 + 1/\kappa)$ [5] (and therefore depends on both κ and β) and the coefficient of variation (CV) is a function of κ alone. Increasing failure rate distributions require $\kappa \geq 1$. An increase of κ decreases the CV, where $\kappa = 1$ corresponds to $CV = 1$ (the exponential distribution). To cover the parameter space, we sampled κ uniformly so that a large range of CVs is covered. More specifically, we have: $\kappa = 1$ ($CV = 1$), $\kappa = 2$ ($CV = 0.523$), $\kappa = 3$ ($CV = 0.363$) and $\kappa = 4$ ($CV = 0.281$). The mean is an increasing function of β and is not very sensitive to values of κ [6]. Choosing $\beta = \{1, 10, 100, 1000\}$ covers scenarios with very low to very high valuation means. The problem parameters, c and ψ, were chosen as a fraction of β. Because an interior solution requires at least $E[V] > c$, and β is a proxy

[3] The proposition applies to a general distribution F and is not restricted to the uniform distribution.

[4] The Weibull distribution was chosen due to its flexibility. When $\kappa = 1$, then the Weibull distribution reduces to the exponential distribution. When $\kappa = 3.4$, then the Weibull distribution mimics the normal distribution. The distribution is IFR (in the weak sense) for $\kappa \geq 1$.

[5] The Gamma function, $\Gamma(z)$, is defined by $\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}dt$. For a positive integer z, $\Gamma(z) = (z-1)!$.

[6] In fact, $E[V] \approx \beta$, where $E[V] = \beta$ for $\kappa = 1$ and $\kappa \to \infty$ and $E[V] < \beta$ for $\kappa \in (1, \infty)$.

Table 2. Parameter values used in the numerical experiment.

Parameter	Values
Valuation Distribution	Weibull(κ, α)
κ	{1, 2, 3, 4}
β	{1, 10, 100, 1000}
ψ	$\{0.001\alpha, 0.01\alpha, 0.05\alpha, 0.1\alpha, 0.5\alpha, 0.75\alpha, \alpha\}$
c	$\{0.001\alpha, 0.01\alpha, 0.05\alpha, 0.1\alpha, 0.5\alpha, 0.75\alpha, \alpha\}$

Table 3. Threshold values $\widetilde{c}$ under which subscription is preferable and over which pay-per-use is preferable at $w/\Lambda = 0.001$ and at $\widehat{w/\Lambda}$ for different values of β and κ of the Weibull distribution.

κ	β	$\widetilde{c}\,(0.001)$	$\widehat{\psi}$	$\widetilde{c}\left(\widehat{\psi}\right)$
1	1	0.841	0.06	0.841
1	10	8.414	0.62	8.414
1	100	84.140	6.93	84.140
2	1	0.752	0.03	0.610
2	10	7.769	0.37	6.031
2	100	78.502	3.74	60.233
3	1	0.781	0.01	0.707
3	10	8.067	0.18	6.723
3	100	81.486	1.89	66.906
4	1	0.808	0.01	0.732
4	10	8.338	0.11	7.274
4	100	84.199	1.10	72.741

for the mean, we can focus on values of c for which $c \leq \beta$. We bound ψ in the same fashion and choose values of c and ψ that partition the range.

In all instances in which the optimal subscription profits were positive – that is, there existed an interior solution – we observed that $v_s < v_t$ and $\mu_s > \mu_t$. That is, as in the uniform distribution, the numerical study suggests that subscription pricing results in over congestion and over investment in capacity.

As was proved for the uniform distribution, we observe that $\widetilde{c}$ is decreasing in ψ for the Weibull distribution as well. In addition, we observe that $\widetilde{c}$ is increasing in β. Keeping all other parameters constant, $\widetilde{c}\,(\kappa)$, however, is not a monotone function. This is not surprising, because the Gamma function is not monotone in κ. Table 3 provides examples of threshold values $\widetilde{c}$ for different values of κ, β and ψ. $\widehat{\psi}$ denotes the maximum value of ψ such that there exists a threshold level $\widetilde{c}$ above which pay-per-use is better. That is, for all values of $\psi > \widehat{\psi}$, selling subscriptions is better than charging on a pay-per-use basis for all values of c. $\widetilde{c}\left(\widehat{\psi}\right)$ is the threshold capacity cost at that maximum value.

7 Third-Party Revenues

The analysis of third-party revenues can be extended to a strictly concave and increasing revenue function in the transaction rate. Strict concavity seems like a more reasonable assumption. Let $r\left(\Lambda\bar{F}(v)\right)$ be total third-party revenue rate the firm incurs if the congestion in the system is $\Lambda\bar{F}(v)$

(where $r'(\cdot) > 0$, $r''(\cdot) < 0$). The per-use profit function is

$$\Pi_p(v_p) = \Lambda \bar{F}(v_p) \left(v_p - \frac{w}{\mu - \Lambda \bar{F}(v_p)} \right) + r\left(\Lambda \bar{F}(v_p)\right)$$

The subscription profit function is:

$$\Pi_s(v_s) = \Lambda \bar{F}(v_s) \left(E\left[V | V \geq v_s\right] - v_s \right) + r\left(\Lambda \bar{F}(v_s)\right)$$

where v_s, as before, is given by

$$v_s = \frac{w}{\mu - \Lambda \bar{F}(v_s)}.$$

Notice that v_s is independent of $r(\cdot)$ whereas $v_p(r)$ is implicitly defined by

$$v_p + r'\left(\Lambda \bar{F}(v_p)\right) = \frac{\bar{F}(v_p)}{f(v_p)} + \frac{w\mu}{\left(\mu - \Lambda \bar{F}(v_p)\right)^2}.$$

The uniqueness of v_p is guaranteed because of the concavity of $r(\cdot)$. Because $r'(\cdot) > 0$, v_p is lower than in the no third-party revenue case. The per-use price charged is lower and more customers request service. As in the fixed revenue per transaction case, it is possible that the optimal per-use price pays customers for the use of the service ($p_p < 0$). If the third-party revenue function is such that

$$\widetilde{r}'\left(\Lambda \bar{F}(v_s)\right) = \frac{\bar{F}(v_s)}{f(v_s)} + \frac{w\Lambda \bar{F}(v_s)}{\left(\mu - \Lambda \bar{F}(v_s)\right)^2}$$

then $v_s = v_p(\widetilde{r})$. In this case, the per-use price is zero, $p_p = 0$, and $\Pi_p(v_p) < \Pi_s(v_s)$.