M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 13, No. 2, Spring 2011, pp. 244–260
ISSN 1523-4614 | EISSN 1526-5498 | 11 | 1302 | 0244
informs
DOI 10.1287/msom.1100.0315

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

Gérard P. Cachon
Operations and Information Management, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, cachon@wharton.upenn.edu

Pnina Feldman
Operations and Information Technology Management, Haas School of Business, University of California, Berkeley, Berkeley, California 94720, feldman@haas.berkeley.edu

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.

Key words: service operations; operations strategy; pricing and revenue management; game theory; queueing theory
History: Received: June 18, 2008; accepted: July 28, 2010. Published online in *Articles in Advance* January 21, 2011.

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: customers pay a per-transaction fee each time they withdraw money from the ATM; beauty shops and hair salons price on a per-use basis; and car maintenance companies, such as Pep Boys Auto, charge each time a service is completed. Naor (1969) began this line of research, and there have 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; et cetera.

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 more 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 two different capacity management scenarios: (i) the firm's service capacity is exogenously fixed; and (ii) the firm endogenously chooses its service capacity in addition to its pricing policy.

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 and 4 compare the two pricing schemes under two different assumptions for how the firm's capacity is determined. Section 5 extends the base model to consider the effects of heterogeneous usage rates. Section 6 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.

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 et al. (2008) do consider additional pricing schemes in queueing models. Randhawa and Kumar (2008) compare per-use pricing with 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 et al. (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. The literature on nonlinear pricing also studies the design of two-part tariffs for congestion-proned services (e.g., Clay et al. 1992, Miravete 1996, Masuda and Whang 2006). They show that when consumers have heterogeneous needs for the service, a menu of two-part tariffs may be optimal. Like Bitran et al. (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. Furthermore, firms might prefer to forgo the additional revenues from using a two-part tariff to save the transaction costs and the administrative burden required for its implementation. Disney, for example, initially charged consumers both to get into the park and for specific rides within the park, but later it abandoned per-ride charges.

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 that the number of ski-lift rides is fixed and fully utilized no matter which 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—resulting in different utilizations of the server. 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 have 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).

2. Model Description

A single firm provides a service to a market with a potential number of M homogeneous customers. Consumers are assumed to be infinitesimal—i.e., every consumer is small relative to the size of the market. 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 a service opportunity occurs, 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, where the support of V is the interval $[0, \bar{v}]$. 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, in contrast to Essegaier et al. (2002). We discuss multiple customer segments in §5.

Although customers value receiving the service, 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).[1] 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. (This definition of a subscription is equivalent to a fixed fee, K, for a finite duration, d, with unlimited usage during that time, where $k = K/d$.) Where useful, we use "p" and "s" subscripts to signify notation associated with the per-use and subscription schemes, respectively.

The server's processing rate is μ. In §3 we assume μ is exogenous, whereas in §4 the firm chooses μ subject to a fee that is proportional to the service rate. $W(\cdot)$ is the expected service time. 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. We assume that $W(\tau M)$ 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 a customer's information about the queue length during one service occasion is of little use in predicting the waiting time for the next service encounter. It also implies that for a fixed potential arrival rate of service, Λ, the potential population of customers, M, is large, they do not seek service too frequently (τ is small), and capacity is sufficiently large that $W(\tau M) \ll 1/\tau$. (For example, the interarrival time of services could be measured in days, whereas service times could be measured in minutes.) Consequently, the arrival rate to the firm's queue does not vary (approximately) with 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 she is in the process of withdrawing cash. (See Randhawa and Kumar 2008 for a model of a closed queueing system in which the arrival rate to the queue depends on the number of customers in queue.) Therefore, the expected service time depends only on the actual arrival rate, λ. The function $W(\lambda)$ is strictly increasing ($W'(\lambda) > 0$) and convex ($W''(\lambda) \geq 0$). (Thus, $W(0) = 1/\mu$ because $1/\mu$

[1] As in Afèche and Mendelson (2004), it is possible to allow the waiting cost to be linear in the value of the service, $w = a + bv$. A detailed analysis is available from the authors.

is a customer's service time when there is no congestion.) Naturally, $W(\lambda)$ is decreasing in μ.

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, $W(\lambda)$.[2] 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 expected service-time cost or the expected congestion cost. Note that 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). 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. We assume that the impact of any one consumer on the average queue length is insignificant. 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.[3]

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 $\kappa > 1$, whereas both (A1) and (A2) hold if F is uniform on the support $[0, \bar{v}]$ or Weibull with parameters $\kappa \geq 1$ and $\beta > 0$. (Note, a Weibull distribution with $\kappa = 1$ is an exponential distribution.) In both versions of the model, we assume F is uniform on the support $[0, \bar{v}]$ to derive analytical comparisons between the pricing schemes. Regarding the system-time function, in the capacity choice model (§4) we assume that $W(\lambda)=1/(\mu-\lambda)$, 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 (§3).

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 nonnegative, i.e., the value of that opportunity is greater than or equal to $p + wW(\lambda)$. 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

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

[2] In fact, the customer only needs 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 functional form of $W(\cdot)$).

[3] 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 endogenous capacity model by the inclusion of mixed strategies. (Refer to §3 of the electronic companion, available at http://msom.pubs.informs.org/ecompanion.html, for a complete analysis of equilibria in mixed strategies under both capacity scenarios.)

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

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

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

That is, the firm's maximization problem can be written as $\max_v R_p(v)$. The following theorem establishes that an optimal threshold, v_p, exists and is unique (proofs are provided in the appendix).

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*

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

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

$$p_p = \frac{\bar{F}(v_p)}{f(v_p)} + w\Lambda \bar{F}(v_p)W'(\Lambda \bar{F}(v_p)). \quad (3)$$

To understand the economic intuition behind (2), note that the first term in the right-hand side (RHS) is the customer's waiting-time cost. The second term is the externality the customer imposes on other consumers due to the (infinitesimal) increase in arrival rate when she decides to join. From a social welfare viewpoint, customers should join as long as their utility from joining (v_p) is larger than the sum of these first two terms. However, the third term, $\bar{F}(v_p)/f(v_p)$, is added by the profit-maximizing monopolist. This correction term implies that the per-use arrival rate in equilibrium is smaller than the social optimal arrival rate. Furthermore, from (3), the optimal per-use fee is greater than the welfare-maximizing per-use fee.

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 that 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 or not to seek service, 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 value of the service opportunity at which a consumer is indifferent, v_s, exactly equals the expected congestion cost:

$$v_s = wW(\Lambda \bar{F}(v_s)). \quad (4)$$

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,

$$\bar{F}(v_s)(E[V \mid V \geq v_s] - v_s),$$

where $\bar{F}(v_s)$ is the probability that a service opportunity is sufficiently valuable to seek service, $E[V \mid V \geq v_s]$ 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 (4)).

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):[4]

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

All consumers purchase a subscription even though they are indifferent between doing so or not, which confirms our initial assumption that all consumers subscribe.[5] As a result, subscription pricing allows the firm to extract all consumer surplus, conditional on the level of congestion that subscriptions generate. The latter condition differentiates this work from the literature on advance selling (e.g., Xie and Shugan 2001)—in those models the potential consumer surplus is independent of the pricing scheme, whereas here it depends on how much congestion materializes.

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

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

Note that although the threshold v_p was a decision variable for the firm with per-use pricing, the

[4] 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.

[5] If mixed strategies are allowed for the consumer purchase decision, then it is possible to show that the firm may be able to earn higher revenue with subscription pricing than we report.

firm does not control the threshold v_s with subscription pricing—it is set by (4). In other words, with exogenous capacity and subscription pricing, the firm cannot control congestion even though it possesses an effective mechanism for maximizing revenue conditional on the system's congestion. That said, congestion is subject to some *self*-regulation—customers request service only for service opportunities whose values exceed their expected congestion cost.

3.3. Comparison Between Per-Use and Subscription Pricing

This section compares the revenues generated by per-use and subscription pricing with an exogenous capacity and quantifies the upper bound on the revenue loss from using these schemes relative to the optimal scheme.

To compare the revenue generated via these two schemes, note that with subscriptions a consumer pays an amount that equals the average net value of her service requests, $E[V \mid V \geq v] - wW(\Lambda\bar{F}(v))$, whereas with per-use pricing, she only pays for the net value of the marginal service request, $v - wW(\Lambda\bar{F}(v))$. Consequently, if the congestion levels were the same with either pricing scheme (i.e., the thresholds v were identical), then subscription pricing clearly generates more revenue. We refer to this as the revenue-extracting benefit of subscription pricing. However, the level of congestion will not be identical (in general) across the two schemes. As one would expect, a comparison of (2) and (4) reveals that per-use pricing results in less congestion (a higher threshold) than subscription pricing: $v_p \geq v_s$. This establishes the trade-off between these two schemes: subscription pricing is better at extracting revenue, but per-use pricing is better at controlling congestion. The firm's preference over these two schemes depends on which of these two effects dominates. For example, in the special case in which consumers are indifferent to congestion (i.e., when $w = 0$), subscription pricing yields higher revenue than per-use pricing and generates more social value per unit time (by having larger usage rates). However, as congestion becomes more costly to consumers (as w increase), per-use pricing may be more attractive.

To make these comparisons more explicit, we assume in the rest of this section that $V \sim U[0, \bar{v}]$ and $W(\lambda) = 1/(\mu - \lambda)$. We now define the set of parameters for which the firm can earn nonnegative revenue. Although the firm's problem is determined by four parameters (w, μ, Λ, and $\bar{v}$), the next theorem indicates that the pricing schemes' relative rankings depend only on two of them.

Lemma 1. *The relative revenue between subscription and per-use pricing, R_s/R_p, can be expressed in terms of α and ρ, where $\alpha = w/\mu\bar{v}$ and $\rho = \Lambda/\mu$, and both revenues are nonnegative for $\alpha \in [0, 1]$.*

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.

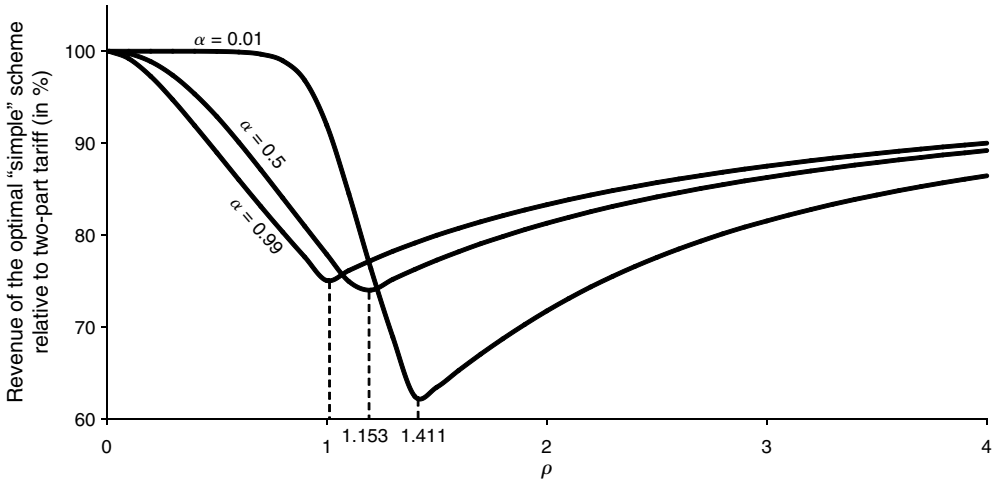
Theorem 2. *When $\alpha = 0$, subscription pricing always yields higher revenue than per-use pricing. For each value of $\alpha > 0$, there exists a unique $\tilde{\rho}(\alpha)$ such that subscription yields higher revenue than per-use pricing for $\rho < \tilde{\rho}(\alpha)$ (recall, ρ is the potential utilization, Λ/μ). Otherwise, per-use pricing yields higher revenue. Moreover, $\tilde{\rho}(\alpha)$ is decreasing in α.*

From Theorem 2, per-use pricing is preferred over subscription for highly congested systems. The key issue is the degree of congestion needed for per-use pricing to be preferred. For various levels of positive congestion costs, $\alpha > 0$, Table 1 provides the potential utilization rate, $\tilde{\rho}(\alpha)$, at which the two schemes yield the same revenue. It can be demonstrated (Proposition 4 of the electronic companion) that $\lim_{\alpha\to 0} \tilde{\rho}(\alpha) = \sqrt{2}$ and $\lim_{\alpha\to 1} \tilde{\rho}(\alpha) = 1$. 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 may also be preferred when the system's actual utilization rate, $\Lambda\bar{F}(v)/\mu$, is high. Table 1 lists the system's actual utilization rate when the potential utilization rate is $\tilde{\rho}(\alpha)$. 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. Of course, subscriptions performs better than per-use pricing under such high utilization rates when the relative congestion cost parameter, α, is low. When α is high, subscriptions will be preferred to per-use pricing if the actual utilizations are more moderate. In addition, it can be shown that the actual utilization rates are increasing in ρ (proof available from authors). 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,

Table 1 **Potential Utilization Rates, $\tilde{\rho}(\alpha)$, That Yield Identical Revenue with Per-Use and Subscription Pricing, as Well as Actual Utilizations When the Potential Arrival Rate Is $\tilde{\rho}(\alpha)\mu$**

		Actual utilization (%) when $\rho = \tilde{\rho}(\alpha)$	
α	$\tilde{\rho}(\alpha)$	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

Figure 1 **Percent of Two-Part Tariff Revenue Attained by the Optimal Simple Scheme as a Function of Potential Utilization, ρ, for Different Values of α**



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 utilization.

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

Subscription pricing generates twice as much revenue as per-use pricing 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 dominate 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.

3.4. Comparison to the Two-Part Tariff

As mentioned in §1, the two-part tariff is the optimal pricing scheme in this setting. A two-part tariff combines a per-use fee with a subscription rate. The firm can set a per-use price to achieve the social optimal congestion, and it can set a subscription price to extract all customer welfare. Hence, social welfare is maximized and is fully extracted by the firm. A complete analysis of the two-part tariff is relegated to Propositions 1–3 of the electronic companion. We characterize the two-part tariff scheme, show that the congestion under two-part tariff is lower than that under subscription pricing and higher than under per-use pricing, and find that the revenue generated by the two-part tariff is well approximated by subscription when ρ is low and by per-use pricing when $\rho \to \infty$.

Despite its ability to extract revenue, as discussed earlier, two-part tariffs may not be offered in practice for reasons that we do not model. Nevertheless, we can evaluate the percent loss of using an optimal "simple" scheme (per-use or subscription) relative to the optimal two-part tariff.[6] Figure 1 illustrates the percent of two-part tariff revenue that is attained by optimally setting one of the two simple schemes as a function of potential utilization, ρ, for different values of α. For any fixed α, the maximum percent revenue loss of the optimal simple scheme relative to the optimal two-part tariff occurs at the potential utilization where both simple schemes are equally profitable (i.e., at $\tilde{\rho}$). The revenue loss is lower at lower values of ρ (subscription pricing is optimal) and higher values of ρ (per-use pricing is optimal), vanishing at $\rho \to 0$ and at $\rho \to \infty$. Moreover, the maximum percent revenue loss is decreasing in α.

4. Capacity Choice

In §3 the firm can choose how to price, but not its capacity, so the pricing decision results only in variation in service time. 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. We assume that capacity is costly—the firm

[6] Giridharan and Mendelson (1994) also quantify the loss of using a suboptimal pricing scheme in a model with congestion.

incurs a cost at rate $c\mu$ for maintaining capacity μ, where $c > 0$. Furthermore, we continue to assume $W(\lambda) = 1/(\mu - \lambda)$.

4.1. 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 value at which consumers are indifferent, v, and capacity, μ:

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

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

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

The firm's profit rate, $\Pi_p(v, \mu_p(v))$, can now be written as

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

where the constant ϕ is defined for convenience:

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

The firm solves the maximization problem, $\max_v \Pi_p(v)$. The following theorem establishes the uniqueness of the optimal per-use threshold.

Theorem 4. *If $\bar{v} > c$, there exists an upper bound $\bar{\phi}_p$ such that for every $\phi < \bar{\phi}_p$ there exists a unique optimal threshold, $v_p = \arg\max_v \Pi_p(v)$, that yields positive profit, $\Pi_p(v_p) > 0$. This threshold is the smallest solution to the implicit equation given by*

$$v_p = \frac{\bar{F}(v_p)}{f(v_p)} + \frac{\phi}{\sqrt{\bar{F}(v_p)}} + c. \tag{5}$$

Furthermore, if (A1) *holds, then there exist two solutions to* (5)*. Otherwise, there does not exist an optimal $v_p < \bar{v}$. The optimal capacity is*

$$\mu_p = \Lambda \bar{F}(v_p) + \sqrt{\frac{w\Lambda \bar{F}(v_p)}{c}}, \tag{6}$$

and the firm's per-use fee is

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

The bound in Theorem 4, $\bar{\phi}_p$, 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. Observe that we restrict attention to the interesting case in which profits are positive. If $\phi > \bar{\phi}_p$ (so that $\Pi_p(v_p) < 0$), the firm is strictly better off not selling per-use, and if $\phi = \bar{\phi}_p$ (so that $\Pi_p(v_p) = 0$), the firm is indifferent.

4.2. 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 $v = w/(\mu - \Lambda \bar{F}(v))$ 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

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

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. The firm solves the maximization problem, $\max_v \Pi_s(v)$.

Theorem 5. *If $E[V] > c$, there exists an upper bound $\bar{\phi}_s$ such that for every $\phi < \bar{\phi}_s$, there exists an optimal threshold, $v_s \in \arg\max \Pi_s(v)$, that yields positive profit, $\Pi_s(v_s) > 0$. This threshold is implicitly defined by*

$$v_s = \phi\sqrt{\frac{1}{\bar{F}(v_s) - cf(v_s)}}. \tag{7}$$

Furthermore, if (A2) *holds, then there exist two solutions to* (7) *and the smallest solution is the unique optimal threshold. The optimal capacity is*

$$\mu_s = \Lambda \bar{F}(v_s) + \frac{w}{v_s}, \tag{8}$$

and the firm's subscription rate is

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

As with Theorem 4, Theorem 5 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, $\bar{\phi}_p$ and $\bar{\phi}_s$, need not be the same. Observe also that as with Theorem 4, we restrict attention to the case in which profits are positive. If $\phi > \bar{\phi}_s$ (so that $\Pi_s(v_s) < 0$), the firm is better off not selling subscriptions, and if $\phi = \bar{\phi}_s$ (so that $\Pi_s(v_s) = 0$), the firm is indifferent.

4.3. Comparison

In this section we assume $V \sim U[0, \bar{v}]$. Let v_t and μ_t be the optimal threshold value and the service rate chosen when the firm uses the two-part tariff scheme. As with fixed capacity, by charging a two-part tariff,

the firm is able to achieve social optimal congestion (by setting an appropriate per-use fee) and to extract all consumer welfare (by setting a subscription rate that makes consumers indifferent to subscribing), so the two-part tariff is optimal for the firm.

As in the fixed-capacity case, we show that per-use pricing leads to a system with less congestion than optimal and that subscription pricing leads to more congestion than optimal: $v_s < v_t < v_p$. Furthermore, the firm invests more in capacity than optimal with subscription pricing (to control congestion somewhat) and less with per-use pricing: $\mu_p < \mu_t < \mu_s$. Even though the firm invests more in capacity with subscription pricing, congestion is also higher with that scheme: $u_s(c) > u_t(c) > u_p(c)$,[7] where $u_x(c)$ is the actual utilization rate,

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

As in the exogenous capacity case, when $w = 0$, congestion is not an issue. The firm's choice is simple. Because there are no congestion costs, the firm has no incentive to provide fast service—it will choose $\mu = \Lambda$, and the capacity cost will be the same for both pricing schemes. That is, similar to the exogenous capacity case, when $w = 0$, subscription pricing dominates per-use pricing for all values of c.

In the discussion that follows, we compare between the two pricing schemes when the marginal congestion cost is positive (i.e., for $w > 0$). If capacity is inexpensive, $c = 0$, subscription pricing performs strictly better than per-use pricing,

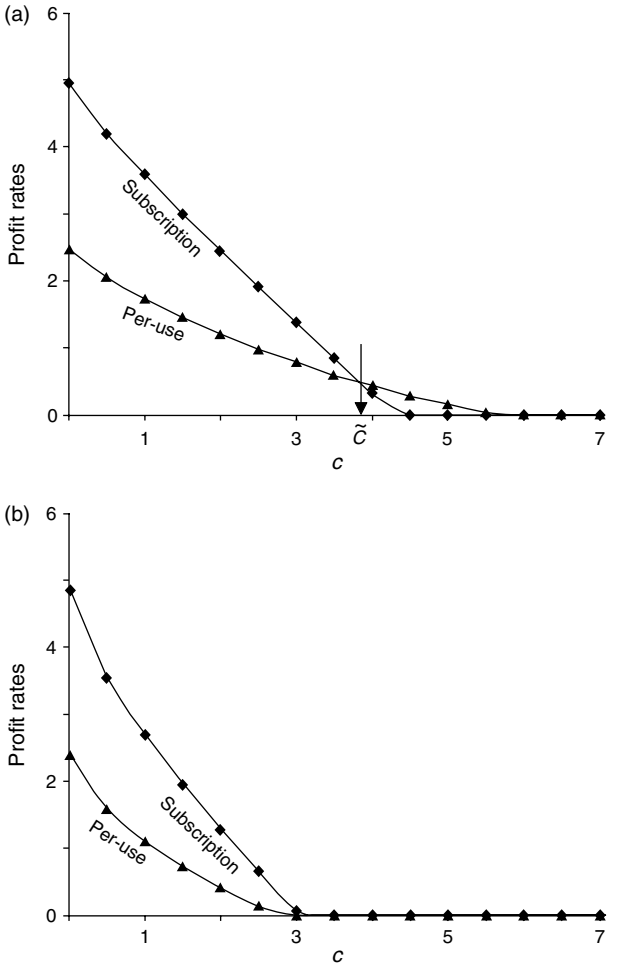
$$\Pi_s(v_s \mid c = 0) > \Pi_p(v_p \mid c = 0);$$

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 nonnegative profit with pricing scheme $x \in \{s, p\}$. Combining these results, one of two scenarios emerges: either $\bar{c}_s \le \bar{c}_p$ or $\bar{c}_p \le \bar{c}_s$.

Figure 2 **Profit Rates of the Two Pricing Schemes with Respect to the Capacity Cost, c**



Note. The following parameter values are used: (a) $w = 0.05$; and (b) $w = 0.5$. ($\Lambda = 1$ and $\bar{v} = 10$ in both panels.)

Figure 2 illustrates these scenarios. On the left hand side, $\bar{c}_s \le \bar{c}_p$. There exists some $\tilde{c}$ such that the two schemes earn the same profit, $\Pi_s(v_s \mid \tilde{c}) = \Pi_p(v_p \mid \tilde{c}) > 0$. It follows that subscription yields higher profit than per-use pricing for $c \in [0, \tilde{c}]$, whereas 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. However, although per-use pricing can be more profitable than subscription pricing, it is only more profitable when capacity is sufficiently expensive, and the absolute advantage of per-use pricing is generally small, whereas the absolute advantage of subscription pricing can be large.

[7] Refer to Propositions 6 and 7 in the electronic companion for formal statements and proofs of these three results. In fact, the results proved in the electronic companion are more general. We are able to show that $v_t < v_p$, $\mu_p < \mu_t$, and $u_t(c) > u_p(c)$ for every IFR distribution. The comparisons of the subscription threshold value with the two-part tariff one was only proven for the uniform distribution, but we observed numerically that the results resported in this section hold for the Weibull distribution with $\kappa \ge 1$ as well. (The cumulative distribution function of a Weibull random variable X is $F(x) = 1 - e^{-(x/\beta)^\kappa}$, where $\beta > 0$ and $\kappa > 0$.) A detailed description of the numerical analysis is available from the authors.

The second scenario, $\bar{c}_p \le \bar{c}_s$, is illustrated on the RHS of Figure 2. 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.[8] (Refer to Proposition 8 in the electronic companion for a formal statement and proof of the result.)

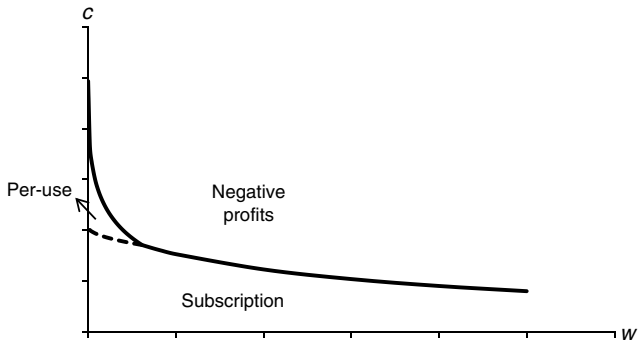
The only difference between the two panels in Figure 2 is that the right-hand side has a higher waiting cost: $w = 0.5$ instead of $w = 0.05$. In fact, it can be shown (Proposition 9 of the electronic companion) that there exists a $\tilde{w}$ such that for all $w > \tilde{w}$ the second scenario occurs, i.e., if the marginal 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. To explain, 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.

Figure 3 illustrates the threshold structure that identifies the regions in which each pricing scheme is more profitable. When the combination of capacity and congestion cost parameters is high (i.e., for (w, c) points located above the solid line), both pricing schemes yield negative profits. Subscriptions are preferred in the lower region and per-use pricing is preferred in the region with low w and high c.

It is also illustrative to compare the pricing schemes with respect to actual utilization. It can be shown that the relevant metric is $w/\Lambda\bar{v}$. (Note that with a fixed capacity we use $w/\mu\bar{v}$ for making comparisons between the two pricing schemes, but now μ is endogenous and different across schemes.) Table 2 provides the firm's actual utilization under each pricing scheme when capacity is $\tilde{c}$, i.e., when the marginal 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 $w > \tilde{w}$ regime.) We observe numerically that actual 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 $w/\Lambda\bar{v} = 0.03$, 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 even if the utilization rate is quite high (say, higher than 98%). Therefore, as in the previous model, subscription pricing can be better than per-use pricing even if it results in a highly utilized system.

[8] This result provides an interesting contrast with the necessary conditions for each pricing scheme to be profitable. Recall that $E[V] > c$ is necessary for subscription pricing, whereas the less restrictive $\bar{v} > c$ is necessary for per-use pricing. These are only necessary conditions, and not sufficient conditions, as we have demonstrated. Therefore, it would be misleading to conclude from those conditions that a high capacity cost favors per-use pricing in all circumstances.

Figure 3 Profitability Regions in $w - c$ Space for Each Pricing Scheme



5. Customers' Heterogeneity in Usage Rates

In our model, customers are heterogeneous in their realized values for service opportunities, but they are otherwise homogeneous. A natural relaxation of this model is to allow heterogeneity in service usage rates. For example, suppose there are two equal-sized segments of consumers. One segment has service opportunities that occur at rate $\tau_h = \tau + \delta$, and the other segment has service opportunities that occur at rate $\tau_l = \tau - \delta$, where $0 \le \delta < \tau$, i.e., τ is the average arrival rate. If consumers do not know a priori which segment they belong to, then our analysis continues to hold because the consumers remain homogeneous in their expectations. To create meaningful segments, it is necessary to assume that consumers know a priori to which segment they belong. This section considers a model with this assumption.

Table 2 Actual Utilization When Capacity Is Such That Subscription Pricing and Per-Use Pricing Yield the Same Profit (i.e., $c = \tilde{c}$)

$w/\Lambda\bar{v}$	0.03	0.02	0.005	0.002	0.0005	0.0001	0.00001
u_p	61.4	67.2	81.9	88.1	93.8	97.2	99.1
u_s	80.2	83.7	91.7	94.8	97.4	98.8	99.6

Interestingly, the per-use results in §§3 and 4 continue to hold without any needed modification. To explain, with per-use pricing each customer makes a decision with each service opportunity, so the rate at which service opportunities occur has no impact on any one decision. All that matters is the aggregate rate at which consumers use the service—as long as τ is held constant, it does not matter if there is one segment ($\delta = 0$) or two segments ($\delta > 0$) or how far apart the two segments are.

On the other hand, the existence of different segments influences subscription pricing. We show that when the service rate is fixed, heterogeneity in usage rates may increase or decrease subscription revenues, but that heterogeneity always decreases profits when the firm chooses the level of capacity. With subscription pricing the firm has two choices. The first is to set the subscription rate such that all types purchase

$$k_l = \tau_l \bar{F}(v_{sl})(E[V \mid V \geq v_{sl}] - v_{sl}),$$

where

$$v_{sl} = wW(M\tau\bar{F}(v_{sl})).$$

The second is to set the subscription rate high enough so that only the high-usage types purchase,

$$k_h = \tau_h \bar{F}(v_{sh})(E[V \mid V \geq v_{sh}] - v_{sh}),$$

where

$$v_{sh} = wW(M\tau_h \bar{F}(v_{sh})/2).$$

Consider first the fixed-capacity model (§3). Let $R_l(\delta)$ and $R_h(\delta)$ be the firm's revenue functions when selling to both segments and to only the high-usage segment, respectively. In our original model, $\delta = 0$, the subscription price is k_l, all consumers subscribe, and the firm extracts all social welfare. As the market becomes more segmented (that is, δ increases) the firm must reduce k_l to capture both segments. The low-type consumers continue to earn zero surplus, but now the high-type consumers are able to retain some surplus, which increases as more segmentation is introduced. Hence, $R_l(0) > R_l(\delta)\ \forall\, \delta \in (0, \tau)$: conditional on all customers subscribing, compared to the original model, subscriptions are less attractive relative to per-use pricing—the increase in δ results in a decrease in subscription revenue, but has no impact on per-use revenue.

This, however, need not be the case if the firm abandons the low segment and prices at k_h to capture only the rent from the high types. By selling only to the heavy users, the firm is able to charge a higher price and extract all social welfare, but it sells to only half of the consumers. These two countervailing effects may result in higher profits relative to the original model.

The next theorem establishes the condition for which heterogeneity in the consumers' usage rates may lead to an increase in the firm's profits under subscription for $V \sim U[0, \bar{v}]$ and $W(\lambda) = 1/(\mu - \lambda)$.

Theorem 6. *$R_h(\delta)$ is quasi-concave. If $\rho > 1$, there exists a unique δ, such that $R_h(\delta) > R_l(0)\ \forall\, \delta \in (\underline{\delta}, \tau)$. Otherwise, $R_h(\delta) \leq R_l(0)\ \forall\, \delta$ and $R_h(\delta)$ is increasing.*

From Theorem 6, if the potential utilization is high, selling subscriptions can become even more attractive when customers are heterogeneous. To explain, remember that although subscription is good at extracting revenues, it cannot control congestion. This becomes more problematic as ρ increases. By selling to only the high-usage customers, the firm forgoes revenue from low-usage consumers, but it is able to control congestion somewhat, while at the same time it extracts the entire customer surplus. In this case, if δ is high enough, this results in higher revenues than the revenues obtained by selling to all consumers at $\delta = 0$.[9]

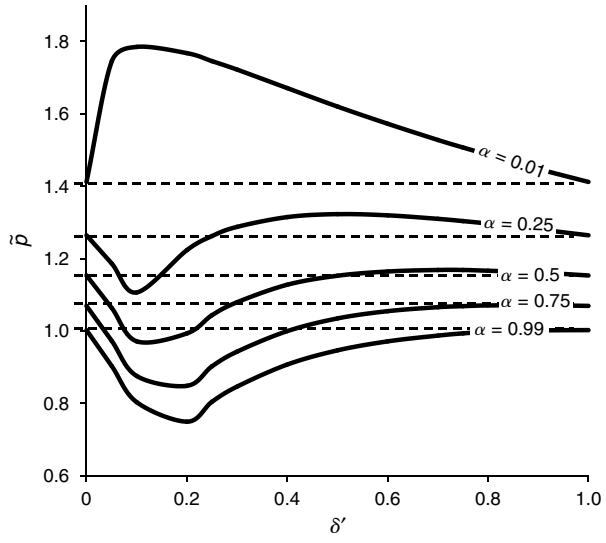
Let $R_s(\delta)$ be the firm's optimal revenue under subscription with degree of heterogeneity δ—i.e., $R_s(\delta) = \max\{R_l(\delta), R_h(\delta)\}$. $R_s(\delta)$ can be fully characterized by α, ρ, and $\delta' = \delta/\tau \in [0, 1]$ (the proof is similar to Lemma 1 and is omitted). The next theorem compares between subscription and per-use revenues when consumers are heterogeneous in usage rates for $V \sim U[0, \bar{v}]$ and $W(\lambda) = 1/(\mu - \lambda)$.

Theorem 7. *When $\alpha = 0$, subscription pricing always yields higher revenue than per-use pricing does. For each value of $\alpha > 0$ and every value of $\delta' \in [0, 1]$, there exists a unique $\tilde{\rho}(\alpha, \delta')$, such that subscription yields higher revenue than per-use pricing for $\rho < \tilde{\rho}(\alpha, \delta')$ (recall, ρ is the potential utilization, Λ/μ and δ' is a standardized measure of heterogeneity, δ/τ). Otherwise, per-use pricing yields higher revenue.*

Theorem 7, which is a generalization of Theorem 2, shows that when consumers are heterogeneous in their usage rates, subscription is preferred over per-use pricing for low levels of ρ and per-use is preferred over subscription for high levels of ρ. It is interesting to assess how the degree of congestion needed for per-use to be preferred changes when consumers are heterogeneous. For various levels of positive congestion costs, $\alpha > 0$, and different degrees of heterogeneity, δ', Figure 4 illustrates the potential utilization rate, $\tilde{\rho}(\alpha, \delta')$, at which the two schemes yield the same revenue. Observe that $\tilde{\rho}(\alpha, 0) = \tilde{\rho}(\alpha, 1)$. This follows because when $\delta' = 0$, consumers are homogeneous and $R_l(0)$ is equivalent to the revenue of the original subscription model. When $\delta' = 1$, the level of heterogeneity is so high that the revenue obtained

[9] This result is analogous to the result obtained by allowing for mixed-strategy equilibrium. Setting a higher subscription price to make only a fraction of consumers join is equivalent to choosing the level of market segmentation, δ. And, in fact, we find that if (and only if) $\rho > 1$, allowing for a mixed-strategy equilibrium benefits subscription even more.

Figure 4 **Potential Utilization Rates, $\tilde{\rho}(\alpha, \delta')$, That Yield Identical Revenue with Per-Use and Subscription Pricing When Consumers Are Heterogeneous**



from selling only to the high-usage consumers—i.e., $R_h(1)$—is equivalent to the revenue of the original subscription model. Observe also that the values of $\tilde{\rho}(\alpha, 0) = \tilde{\rho}(\alpha, 1)$ (marked by the dashed lines in Figure 4) correspond to the values presented in Table 1. Figure 4 demonstrates that the degree of congestion needed for per-use pricing to be preferred over subscription when consumers are heterogeneous can either increase or decrease. This result follows from the two countervailing effects that heterogeneity has on subscription revenues: when consumers are heterogeneous, the firm cannot extract all revenues by selling to all consumers, and thus the revenue accrued by selling to all consumers decreases. However, when the potential utilization is high, the firm can benefit from selling to only the high-usage consumers if the degree of heterogeneity is high (Theorem 6).

When capacity choice is endogenous (§4), the firm that sells subscriptions can control congestion not only by setting a high subscription price so that only high-type consumers purchase, but also by choosing the service rate. As in §4.2, instead of solving for μ, we can frame the firm's problem in terms of choosing the thresholds v_l (if the firm decides to set a subscription price so that all consumers purchase) and v_h (if the firm sets a high subscription price so that only high-usage consumers purchase). Consider first the choice of capacity conditional on all consumers subscribing. In this case,

$$\mu_l = \Lambda \bar{F}(v_l) + w/v_l$$

and the firm's profit function can be written as

$$\Pi_l(v_l; \delta) = M\tau_l \bar{F}(v_l)(E[V \mid V \geq v_l] - v_l) - c(\Lambda \bar{F}(v_l) + w/v_l).$$

The firm solves the maximization problem, $\max_{v_l} \Pi_l(v_l; \delta)$.

THEOREM 8. *If $(\tau_l/\tau)E[V] > c$, there exists an upper bound $\bar{\phi}_l$, such that for every $\phi \leq \bar{\phi}_l$, there exists an optimal threshold $v_{sl} \in \arg\max \Pi_l(v_l; \delta)$, that yields nonnegative profit, $\Pi_l(v_{sl}; \delta) \geq 0$. This threshold is implicitly defined by*

$$v_{sl} = \phi \sqrt{\frac{1}{(\tau_l/\tau)\bar{F}(v_{sl}) - cf(v_{sl})}}. \tag{9}$$

Furthermore, if (A2) *holds, then there exist two solutions to* (9) *and the smallest solution is the unique optimal threshold conditional on all consumers subscribing.*

If, however, the firm decides to cater only to the high-usage consumers,

$$\mu_h = M\tau_h \bar{F}(v_h)/2 + w/v_h,$$

and the firm's profit function is

$$\Pi_h(v_h; \delta) = M\tau_h \bar{F}(v_h)(E[V \mid V \geq v_h] - v_h)/2 - c(M\tau_h \bar{F}(v_h)/2 + w/v_h).$$

The firm solves the maximization problem, $\max_{v_h} \Pi_h(v_h; \delta)$.

THEOREM 9. *If $E[V] > c$, there exists an upper bound $\bar{\phi}_h$, such that for every $\phi \leq \bar{\phi}_h$, there exists an optimal threshold $v_{sh} \in \arg\max \Pi_h(v_h; \delta)$ that yields nonnegative profit, $\Pi_h(v_{sh}; \delta) \geq 0$. This threshold is implicitly defined by*

$$v_{sh} = \phi \sqrt{\frac{\Lambda}{M\tau_h/2} \cdot \frac{1}{\bar{F}(v_{sh}) - cf(v_{sh})}}. \tag{10}$$

Furthermore, if (A2) *holds, then there exist two solutions to* (10) *and the smallest solution is the unique optimal threshold conditional on only high-usage consumers subscribing.*

Finally, the firm sets the subscription price that yields maximum profits, $\max\{\Pi_l(v_{sl}; \delta), \Pi_h(v_{sh}; \delta)\}$. We show that when the firm chooses the level of capacity, customer heterogeneity *always* decreases profits under subscription (Proposition 11 of the electronic companion). In this case the firm is able to control congestion by choosing the service rate, μ. Whereas selling only to heavy users had a congestion control advantage in the fixed-capacity case, here the firm loses from selling to fewer consumers. Hence, compared to the original model, subscriptions are less attractive, in a relative sense, than per-use pricing. Moreover, the threshold capacity cost under which subscription profits are higher than per-use profits is lower than in the original model (Corollary 1 of the electronic companion).

We conclude that customer heterogeneity in usage rates generally makes per-use pricing more attractive relative to subscription pricing. However, it can make subscription pricing even more attractive if capacity is fixed and ρ and δ are high.

Finally, we suspect that with heterogeneity in usage rates, a single two-part tariff is generally no longer optimal for the firm—given that there are multiple segments, a single contract that is designed to serve both types will no longer be able to extract the entire surplus. In this situation the firm will typically improve its profits by designing a menu of two-part tariffs or nonlinear tariffs (e.g., Clay et al. 1992, Miravete 1996, Masuda and Whang 2006).

6. 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 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 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 higher profit only when revenue or profit is reasonably low. If customers are heterogeneous in their service rates, subscription pricing can become even more attractive if capacity is fixed and the potential utilization and market segmentation are high. Otherwise, heterogeneity in usage rates makes subscription pricing less attractive relative to per-use pricing. 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.

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

Acknowledgments

The authors thank Albert Ha, the seminar participants at Hong Kong University of Science and Technology, and conference attendees at the M&SOM conference in Maryland, the INFORMS Revenue Management and Pricing Section Conference in Montreal, and the INFORMS Annual Meeting in Washington, DC.

Appendix. Proofs

Let $g(\alpha, \rho) \equiv v_s/\bar{v}$ and $l(\alpha, \rho) \equiv v_p/\bar{v}$. The functions g and l will be used in the proofs of Lemma 1 and Theorem 2 below.

Proof of Theorem 1. Let $\varrho(v) = dR_p(v)/dv$, where

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

There is at least one maximum because $\varrho(0) > \Lambda$ and $\lim_{v\to\infty} \varrho(v) \le 0$. By construction, v_p, given by (2), satisfies the first-order condition. Take $\varrho(v_p) = 0$ and rearrange terms

$$1 - \frac{\bar{F}(v_p)}{v_p f(v_p)} = \frac{wW(\Lambda\bar{F}(v_p))}{v_p} + \frac{w\Lambda\bar{F}(v_p)W'(\Lambda\bar{F}(v_p))}{v_p}. \quad (11)$$

The RHS of (11) is decreasing in v_p because $\bar{F}(v) = \Pr\{V \ge v\}$ is decreasing in v and $W(\cdot)$ is increasing and convex. The left-hand side (LHS) of (11) is increasing in v_p because F is IFR. (Actually, the left-hand side is increasing even if F has an increasing generalized failure rate.) Thus, there is a unique v_p that satisfies $\varrho(v_p) = 0$. □

Proof of Lemma 1. For the per-use case, Theorem 1 establishes that there is a unique optimal v, which is $\bar{v}$ when $R_p'(v_p) \ge 0$. That condition simplifies to $w/\Lambda \ge \bar{v}/\rho$, or $\alpha \ge 1$. It follows that $R_p(v_p) \ge 0$ for all $\alpha \in [0, 1]$. With subscription pricing, revenue is $R_s(v_s) = \Lambda\bar{F}(v_s)(E[V \mid V \ge v_s] - v_s)$. For $V \sim U[0, \bar{v}]$, positive revenues occur $\forall\, v_s < \bar{v}$. Note that the threshold v_s is given by $v_s = w/(\mu - \Lambda\bar{F}(v_s))$, where the LHS is increasing and the RHS is decreasing. Thus, for a nonnegative revenue to occur, we must have that (when evaluating the above condition), at $\bar{v}$, $\bar{v} - w/\mu \ge 0$, or $\alpha \le 1$. Next, we show that the relative revenues are a function of α and ρ. For the uniform distribution, condition (4) can be written as

$$\frac{v_s}{\bar{v}} = \frac{\alpha}{1 - \rho(1 - v_s/\bar{v})}. \quad (12)$$

Note that g is a function of α and ρ only. Plugging v_s into the subscription revenue function, we obtain $R_s = ((1 - g(\alpha, \rho))^2/2)\Lambda\bar{v}$. Similarly, for the pay-per-use case, we can express condition (2) as

$$\frac{v_p}{\bar{v}} = \frac{1}{2}\left(1 + \frac{\alpha}{(1 - \rho(1 - v_p/\bar{v}))^2}\right). \quad (13)$$

Note that l is a function of α and ρ only. Plugging v_p into the per-use revenue function, we obtain

$$R_p = (1 - l(\alpha, \rho))\left(l(\alpha, \rho) - \frac{\alpha}{1 - \rho(1 - l(\alpha, \rho))}\right)\Lambda\bar{v}.$$

Thus, R_s/R_p depends only on α and ρ. □

Proof of Theorem 2. The fact that subscription always does better than per-use pricing when $w = 0$ is immediate. In that case, $\Lambda\bar{F}(v_s)E[V \mid V \ge v_s] \ge \Lambda\bar{F}(v_p)E[V \mid V \ge v_p] > \Lambda\bar{F}(v_p)v_p$, where the first term is the revenue generated by subscription pricing and the last term is the revenue generated by per-use. Note that the first inequality follows because $v_s \le v_p$. The remainder of the proof establishes the result for $w > 0$. Uniqueness: by implicitly differentiating the revenue functions, we get

$$\frac{\partial R_s(\rho)}{\partial \rho} = -\frac{\alpha(1 - g)^2}{\alpha\rho + (1 - \rho(1 - g))^2} \cdot \Lambda\bar{v} \quad (14)$$

and

$$\frac{\partial R_p(\rho)}{\partial \rho} = -\frac{\alpha(1-l)^2}{(1-\rho(1-l))^2} \cdot \Lambda \bar{v}, \tag{15}$$

where g and l are shorthand notation for $g(\alpha, \rho)$ and $l(\alpha, \rho)$. To show that there exists a unique $\tilde{\rho}$ such that $R_s(\rho, \delta') > R_p(\rho)\ \forall \rho < \tilde{\rho}$ and $R_s(\rho) < R_p(\rho)\ \forall \rho > \tilde{\rho}$, it is enough to require that $\forall \rho$ for which $R_s(\rho) > R_p(\rho)$, $R_s'(\rho) < R_p'(\rho)$. $R_s(\rho) > R_p(\rho)$ implies that

$$\frac{(1-g)^2}{2(1-l)} > l - \frac{\alpha}{1-\rho(1-l)}.$$

Rearranging (14) and (15), requiring $R_s'(\rho) < R_p'(\rho)$, we must have that

$$\frac{(1-g)^2}{2(1-l)} > \frac{(1-l)(\alpha\rho + (1-\rho(1-g))^2)}{2(1-\rho(1-l))^2}.$$

Thus, to complete the proof, it is enough to show that

$$l - \frac{\alpha}{1-\rho(1-l)} > \frac{(1-l)(\alpha\rho + (1-\rho(1-g))^2)}{2(1-\rho(1-l))^2}. \tag{16}$$

Plugging in l (from condition (13)) into the LHS of (16) and rearranging, condition (16) becomes

$$\frac{\alpha}{1-\rho(1-l)} + (1-l)\left(\frac{1-\rho(1-g)}{1-\rho(1-l)}\right)^2 < 1. \tag{17}$$

To see that condition (17) holds, note that the first term is smaller than l (follows from the fact that $v_s(\rho) < v_p(\rho)\ \forall \rho$, which implies that the term is less than g and that $g < l$) and that the second term is less than $(1-l)$ (because $g < l$ implies that the squared term is less than 1).

The threshold utilization factor, $\tilde{\rho}$, is implicitly defined by

$$\mathcal{F} = \frac{(1-g(\alpha, \tilde{\rho}))^2}{2} - (1-l(\alpha, \tilde{\rho})) \cdot \left(l(\alpha, \tilde{\rho}) - \frac{\alpha}{1-\tilde{\rho}(1-l(\alpha, \tilde{\rho}))}\right) = 0. \tag{18}$$

Differentiating $\mathcal{F}$ with respect to α yields

$$\frac{d\mathcal{F}}{d\alpha} = \frac{\partial \mathcal{F}}{\partial \alpha} + \frac{\partial \mathcal{F}}{\partial \tilde{\rho}} \cdot \frac{\partial \tilde{\rho}}{\partial \alpha} + \frac{\partial \mathcal{F}}{\partial g} \cdot \frac{\partial g}{\partial \alpha} + \frac{\partial \mathcal{F}}{\partial l} \cdot \frac{\partial l}{\partial \alpha} = 0.$$

Note that the last term equals zero (from the first-order condition). Rearranging the terms, we find that

$$\frac{\partial \tilde{\rho}}{\partial \alpha} = -\frac{(\partial \mathcal{F}/\partial \alpha + \partial \mathcal{F}/\partial g) \cdot \partial g/\partial \alpha}{\partial \mathcal{F}/\partial \tilde{\rho}} = -\frac{(1-l)/(1-\tilde{\rho}(1-l)) + (1-g)(\partial g/\partial \alpha)}{\alpha(1-l)^2/(1-\tilde{\rho}(1-l))^2},$$

where

$$\frac{\partial g}{\partial \alpha} = \frac{1}{1-\rho(1-g)+\rho g} > 0.$$

Thus, we then get that $\partial \tilde{\rho}(\alpha)/\partial \alpha < 0$. Existence: it was established in Proposition 4 of the technical appendix that a threshold $\tilde{\rho}$ exists for $\alpha = 0$ and $\alpha = 1$. Because $\partial \tilde{\rho}(\alpha)/\partial \alpha < 0$, existence is guaranteed $\forall \alpha \in [0, 1]$. □

Proof of Theorem 3. (i) $\rho = 0$: Substituting $\rho = 0$ in Equations (12) and (13) results in $v_s = \alpha \bar{v}$ and $v_p = (1+\alpha)\bar{v}/2$. Then, the following expressions for the revenue rates are immediate:

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

(ii) $\rho \to \infty$: Rearranging (12), we get

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

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

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

Similarly, rewriting (13), we get

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

This implies that for all pricing schemes,

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

Substituting into the revenue rates, we obtain for $\rho \to \infty$: $\lim_{\rho\to\infty} R_s = \lim_{\rho\to\infty} R_p = 0$. □

Proof of Theorem 4. 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 nonnegative profits. Our conditions imply for both $v \in \{v_1, v_3\}$ that

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

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

where the first condition is the first-order condition and the second condition ensures nonnegativity of profits. Combining the two conditions, we have

$$\frac{\bar{F}(v)}{f(v)} \geq \frac{\phi}{\sqrt{\bar{F}(v)}}. \tag{19}$$

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 (19), $\Pi_p'(v_2) = 0$ and $\Pi_p(v_2) < 0$ imply

$$\frac{\bar{F}(v_2)}{f(v_2)} < \frac{\phi}{\sqrt{\bar{F}(v_2)}}. \tag{20}$$

Because F is IFR, the LHS of (20) is decreasing. Furthermore, the RHS of (20) is increasing. As a result, $v_1 < v_2 < v_3$ implies

$$\frac{\bar{F}(v_1)}{f(v_1)} > \frac{\bar{F}(v_2)}{f(v_2)} > \frac{\bar{F}(v_3)}{f(v_3)} \tag{21}$$

and

$$\frac{\phi}{\sqrt{\bar{F}(v_3)}} > \frac{\phi}{\sqrt{\bar{F}(v_2)}} > \frac{\phi}{\sqrt{\bar{F}(v_1)}}. \tag{22}$$

Combining (20), (21), and (22) 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 (19).

Consider the second case, $\Pi_p(v_2) > 0$. Rearranging the first-order condition, let

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

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 positive. Equation (19) implies that 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. Let $(0, \bar{v})$ be the support of $F(v)$, then $\lim_{v\to\bar{v}} \Pi_p(v) = 0$. 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 $\partial M_p(\phi)/\partial\phi = -\Lambda\sqrt{\bar{F}(v_p)} < 0$, which means that $\Pi_p(v_p(\phi), \phi)$ is decreasing in ϕ. This implies that there exists some $\bar{\phi}_p$ such that $\Pi_p(v_p(\phi), \phi) > 0$ for $\phi \leq \bar{\phi}_p$. Otherwise, there does not exist an optimal $v_p < \bar{v}$.

The smallest solution to (5) is v_p; although we cannot compute the number of possible solutions to (5) for a general IFR distribution F, we show by contradiction that the optimal v_p is the smallest solution to (5). Note first that $\Pi_p(0) = -\Lambda(c + 2\phi) < 0$ and that $\Pi_p'(0) > 0$. Thus, the smallest solution to (5) is a local maximum. We have already shown that if there exists a local maximum so that $\Pi_p(v) \geq 0$, it is unique. Suppose there exist two local maxima with negative profits v_1 and v_3 such that $v_1 < v_3$, i.e., $\Pi_p(v_1) < 0$, $\Pi_p(v_3) < 0$, $\Pi_p'(v_1) = 0$, $\Pi_p'(v_3) = 0$. Our conditions for both $v \in \{v_1, v_3\}$ imply that

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

$$\frac{v - c}{2} < \frac{\phi}{\sqrt{\bar{F}(v)}}.$$

Combining the two conditions, we have

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

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

$$\frac{\bar{F}(v_2)}{f(v_2)} > \frac{\phi}{\sqrt{\bar{F}(v_2)}}. \tag{23}$$

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

$$\frac{\bar{F}(v_1)}{f(v_1)} > \frac{\bar{F}(v_2)}{f(v_2)} \tag{24}$$

and

$$\frac{\phi}{\sqrt{\bar{F}(v_2)}} > \frac{\phi}{\sqrt{\bar{F}(v_1)}}. \tag{25}$$

Combining conditions (23) and (25), 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 (24). Letting $v_2 < v_3$, however, a contradiction cannot be reached, which does not preclude the existence of additional solutions to (5) in the negative range for a general IFR distribution. Combining the results, we conclude that v_p is the smallest v that solves (5).

Sufficient condition for at most two solutions to (5): Rearranging (5), 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 $1 + h'(v_p)/(h(v_p))^2$. Thus, if condition (A1) holds, there can be at most two solutions to (5), with the smallest one being the maximum. □

Proof of Theorem 5. First note that $\Pi_s(0) = -\infty$ and that $\lim_{v\to\bar{v}} \Pi_s(v) = 0$. 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 (7) 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

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

for some v. Assume $\phi = 0$. Then, if $E[V] > c$, there must be a solution with positive profit. Let $M_s(\phi) \equiv \Pi_s(v_s(\phi), \phi)$. From the Envelope Theorem, we have $\partial M_s(\phi)/\partial\phi = -2\phi\Lambda/v_s < 0$, which means that $\Pi_s(v_s(\phi), \phi)$ is decreasing in ϕ. Note that even though we have not ruled out the existence of several local maxima v_s, $\Pi_s(v_s(\phi), \phi)$ is decreasing in ϕ at every critical point. This implies that there exists some $\bar{\phi}_s$ such that $\Pi_s(v_s(\phi), \phi)$ for $\phi \leq \bar{\phi}_s$. Otherwise, there does not exist an optimal $v_s < \bar{v}$.

Furthermore, denote the RHS of (7) by $z(v)$, i.e., $z(v) = \phi(1 - c \cdot h(v))^{-1/2}(\bar{F}(v))^{-1/2}$. We want to show that there exists a unique v_s that maximizes profit and solves

$v_s = z(v_s)$. Because F is IFR, $z(v)$ is increasing. Differentiating $z(v)$, we get

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

Plugging in (7), we get

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

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 $h'(v_s)v_s$ is increasing and F is IFR. Thus, under these conditions, $z(v_s)$ is increasing and convex and there are at most two solutions to $v = z(v)$. Because $\Pi_s(0) < 0$ and $\Pi_s'(0) > 0$, the smallest solution is the maximum. □

Proof of Theorem 6. First note that $R_h(\tau) = R_l(0)$. Existence: differentiating $R_h(\delta)$ with respect to δ, we get

$$\frac{dR_h(\delta, v_{sh}(\delta))}{d\delta} = \frac{\partial R_h(\delta, v_{sh}(\delta))}{\partial \delta} + \frac{\partial R_h(\delta, v_{sh}(\delta))}{\partial v_s} \cdot \frac{dv_{sh}}{d\delta} = \frac{M(\bar{v} - v_{sh})^2}{4\bar{v}} - \frac{M\tau_h(\delta)(\bar{v} - v_{sh})}{2\bar{v}} \cdot \frac{dv_{sh}}{d\delta},$$

where $dv_{sh}(\delta)/d\delta > 0$. Let $\varrho(\delta) = dR_h(\delta)/d\delta$. Because $\lim_{\delta \to 0} \varrho(\delta) > 0$ and the domain of δ is bounded above, there exists at least one maximum. Uniqueness: rewrite $v_{sh} = wW(M\tau_h\bar{F}(v_{sh})/2)$ in terms of δ:

$$\delta(v_{sh}) = \frac{2(\mu - w/v_{sh})}{M\bar{F}(v_{sh})} - \tau. \tag{26}$$

Plugging the expression in the revenue function, we get

$$R_h(v_{sh}) = \left(\mu - \frac{w}{v_{sh}}\right)(E[V \mid V \geq v_{sh}] - v_{sh}) = \left(\mu - \frac{w}{v_{sh}}\right)\left(\frac{\bar{v} - v_{sh}}{2}\right),$$

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

$$v_{sh} = \sqrt{\frac{w\bar{v}}{\mu}}, \tag{27}$$

which is unique. Thus, $R_h(\delta)$ is quasi-concave. Plugging Equation (27) in (26) and simplifying, we obtain that $\tau + \delta(v_s) = 2\mu/M$. This δ is achievable if $\Lambda > \mu$, and the results follow. Otherwise, the maximum $R_h(\delta)$ is obtained for $\delta = \tau$ and $R_l(0) > R_h(\delta)\ \forall\, \delta \in (0, \tau)$. □

Proof of Theorem 7. By implicitly differentiating the revenue functions, we get

$$\frac{\partial R_l(\rho, \delta')}{\partial \rho} = -\frac{\alpha(1 - \delta')(1 - g_l)^2}{\alpha\rho + (1 - \rho(1 - g_l))^2} \cdot \Lambda\bar{v}, \tag{28}$$

$$\frac{\partial R_h(\rho, \delta')}{\partial \rho} = -\frac{\alpha((1+\delta')/2)^2(1 - g_h)^2}{\alpha((1+\delta')/2)\rho + (1 - ((1+\delta')/2)\rho(1 - g_h))^2} \cdot \Lambda\bar{v} \tag{29}$$

and $\partial R_p(\rho)/\partial\rho$, given by (15), where $g_l \equiv v_{sl}/\bar{v}$ and $g_h \equiv v_{sh}/\bar{v}$ (analogously to g and l). To complete the proof, it is enough to show that (1) there exists at most one $\tilde{\rho}$, such that $R_l(\rho, \delta') > R_p(\rho)\ \forall\, \rho < \tilde{\rho}$ and $R_l(\rho, \delta') < R_p(\rho)\ \forall\, \rho > \tilde{\rho}$; (2) there exists a unique $\tilde{\rho}$, such that $R_h(\rho, \delta') > R_p(\rho)\ \forall\, \rho < \tilde{\rho}$ and $R_h(\rho, \delta') < R_p(\rho)\ \forall\, \rho > \tilde{\rho}$; and (3) $R_l(\rho = 0, \delta') < R_h(\rho = 0)$, then $R_l(\rho, \delta') < R_h(\rho, \delta')\ \forall\rho$.

(1) $R_l(\rho, \delta') > R_p(\rho)$ implies that

$$\frac{(1 - \delta')(1 - g_l)^2}{2(1 - l)} > l - \frac{\alpha}{1 - \rho(1 - l)}.$$

Rearranging (28) and (15), requiring $R_l'(\rho, \delta') < R_p'(\rho)$, we must have that

$$\frac{(1 - \delta')(1 - g)^2}{2(1 - l)} > \frac{(1 - l)(\alpha\rho + (1 - \rho(1 - g))^2)}{2(1 - \rho(1 - l))^2}.$$

Thus, to complete the proof, it is enough to show that

$$l - \frac{\alpha}{1 - \rho(1 - l)} > \frac{(1 - l)(\alpha\rho + (1 - \rho(1 - g))^2)}{2(1 - \rho(1 - l))^2},$$

which follows from the proof of Theorem 2.

(2) Uniqueness: $R_h(\rho, \delta') > R_p(\rho)$ implies that

$$\frac{((1+\delta')/2)(1 - g_h)^2}{2(1 - l)} > l - \frac{\alpha}{1 - \rho(1 - l)}.$$

Rearranging (29) and (15), requiring $R_h'(\rho, \delta') < R_p'(\rho)$, we must have that

$$\frac{((1+\delta')/2)(1 - g_h)^2}{2(1 - l)} > \frac{(1 - l)(\alpha(1+\delta')\rho + (1 - \rho((1+\delta')/2)(1 - g_h))^2)}{2(1 - \rho(1 - l))^2}.$$

Thus, to complete the proof, it is enough to show that

$$l - \frac{\alpha}{1 - \rho(1 - l)} > \frac{(1 - l)(\alpha(1+\delta')\rho + (1 - \rho((1+\delta')/2)(1 - g_h))^2)}{2(1 - \rho(1 - l))^2}. \tag{30}$$

Plugging in l (from condition (13)) into the LHS of (30) and rearranging, condition (30) becomes

$$\frac{\alpha}{(1 - \rho(1 - l))^2} + (1 - l)\left(\frac{1 - \rho((1+\delta')/2)(1 - g_h)}{1 - \rho(1 - l)}\right)^2 - \frac{\alpha\rho(1 - l)(1 - \delta')}{(1 - \rho(1 - l))^2} < 1. \tag{31}$$

To see that condition (31) holds, note that the first term is smaller than l (follows from (13)), that the second term is smaller than $(1 - l)$ (because $g_h < l$ implies that the squared term is less than 1), and the third term is negative. Existence: it is established in Proposition 10 of the technical appendix that $\forall\delta'$ a threshold $\tilde{\rho}_h(\alpha; \delta')$ for which $R_h(\tilde{\rho}_h, \delta') = R_p(\tilde{\rho}_h)$ exists for $\alpha = 0$ and $\alpha = 1$. Because $\partial\tilde{\rho}_h(\alpha; \delta')/\partial\alpha < 0$ (proof is similar to that of Theorem 2 and is therefore omitted), existence is guaranteed $\forall\alpha \in [0, 1]$.

(3) Note first that $g_l = g$ and $g_h \leq g$. Therefore, $g_h \leq g_l$. $R_l(\rho = 0, \delta') < R_h(\rho = 0)$ if and only if $\delta' > 1/3$. Thus, we need to show that $R_l(\rho, \delta') < R_h(\rho, \delta')\ \forall\rho$ and $\forall\delta' > 1/3$. Substituting the revenue functions and rearranging, it is enough to show that

$$\frac{1 + \delta'}{2}(1 - g_h)^2 > (1 - \delta')(1 - g_l)^2.$$

Because $g_h \leq g_l$, it is enough to show that $(1 + \delta')/2 > 1 - \delta'$, which holds for $\delta' > 1/3$. □

Proof of Theorem 8. First note that $\Pi_l(0)=-\infty$ and that $\lim_{v_l\to\bar{v}}\Pi_l(v_l)=0$. Differentiating $\Pi_l(v_l)$, we obtain

$$\frac{d\Pi_l(v_l)}{dv_l}=c\left(\frac{w}{v_l^2}+\Lambda f(v_l)\right)-M\tau_l\bar{F}(v_l).$$

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

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

$$\frac{\Pi_l(v_l)}{\Lambda\bar{F}(v_l)}=\frac{\tau_l}{\tau}(E[V\mid V\geq v_l]-v_l)-c-\frac{\phi^2}{v_l\bar{F}(v_l)}\geq 0$$

for some v_l. Assume $\phi=0$. Then, if $(\tau_l/\tau)E[V]>c$, there must be a solution with positive profit. Let $M_l(\phi)\equiv\Pi_l(v_{sl}(\phi),\phi)$. From the Envelope Theorem, we have $\partial M_l(\phi)/\partial\phi=-2\phi\Lambda/v_{sl}<0$, which means that $\Pi_l(v_{sl}(\phi),\phi)$ is decreasing in ϕ. Note that even though we have not ruled out the existence of several local maxima v_{sl}, $\Pi_l(v_{sl}(\phi),\phi)$ is decreasing in ϕ at every critical point. This implies that there exists some $\bar{\phi}_l$ such that $\Pi_l(v_{sl}(\phi),\phi)$ for $\phi\leq\bar{\phi}_l$. Otherwise, there does not exist an optimal $v_{sl}<\bar{v}$.

Furthermore, denote the RHS of (9) by $z(v_l)$, i.e., $z(v_l)=\phi((\tau_l/\tau)-c\cdot h(v_l))^{-1/2}(\bar{F}(v_l))^{-1/2}$. We want to show that there exists a unique v_{sl} that maximizes profit and solves $v_{sl}=z(v_{sl})$. Because F is IFR, $z(v_l)$ is increasing. Differentiating $z(v_l)$, we get

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

Plugging in (9), we get

$$z'(v_{sl})=\frac{1}{2}\left(h(v_{sl})v_{sl}+\frac{c\cdot h'(v_{sl})v_{sl}}{(\tau_l/\tau)-c\cdot h(v_{sl})}\right).$$

A sufficient condition for $z'(v_{sl})$ 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 $h'(v_{sl})v_{sl}$ is increasing and F is IFR. Thus, under these conditions, $z(v_{sl})$ is increasing and convex and there are at most two solutions to $v_l=z(v_l)$. Because $\Pi_l(0)<0$ and $\Pi_l'(0)>0$, the smallest solution is the maximum. □

Proof of Theorem 9.First note that $\Pi_h(0)=-\infty$ and that $\lim_{v_h\to\bar{v}}\Pi_h(v_h)=0$. Differentiating $\Pi_h(v_h)$, we obtain

$$\frac{d\Pi_h(v_h)}{dv_h}=c\left(\frac{w}{v_h^2}+M\tau_h f(v_h)/2\right)-M\tau_h\bar{F}(v_h)/2.$$

Equating to zero and rearranging terms, the result in (10) follows.

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

$$\frac{\Pi_h(v_h)}{M\tau_h\bar{F}(v_h)/2}=E[V\mid V\geq v_h]-v_h-c-\frac{\Lambda}{M\tau_h/2}\frac{\phi^2}{v_h\bar{F}(v_h)}\geq 0$$

for some v_h. Assume $\phi=0$. Then, if $E[V]>c$, there must be a solution with positive profit. Let $M_h(\phi)\equiv\Pi_h(v_{sh}(\phi),\phi)$. From the Envelope Theorem, we have $\partial M_h(\phi)/\partial\phi=-2\phi\Lambda/v_{sh}<0$, which means that $\Pi_h(v_{sh}(\phi),\phi)$ is decreasing in ϕ. Note that even though we have not ruled out the existence of several local maxima v_{sh}, $\Pi_h(v_{sh}(\phi),\phi)$ is decreasing in ϕ at every critical point. This implies that there exists some $\bar{\phi}_h$ such that $\Pi_h(v_{sh}(\phi),\phi)$ for $\phi\leq\bar{\phi}_h$. Otherwise, there does not exist an optimal $v_{sh}<\bar{v}$.

Furthermore, denote the RHS of (10) by $z(v_h)$, i.e., $z(v_h)=\phi\sqrt{\Lambda/(M\tau_h/2)}(1-c\cdot h(v_h))^{-1/2}(\bar{F}(v_h))^{-1/2}$. Note that $z(v_h)=z(v)\cdot\kappa$, where $z(v)$ is the RHS of (7) from Theorem 5 and κ is a positive constant. Therefore, the remainder of the proof is the same. □

References

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

Barro, R. J., P. M. Romer. 1987. Ski-lift pricing, with applications to labor and other markets. *Amer. Econom. Rev.* **77**(5) 875–890.

Berglas, E. 1976. On the theory of clubs. *Amer. Econom. Rev.* **66**(2) 116–121.

Bitran, G., P. Rocha e Oliveira, A. Schilkrut. 2008. Managing customer relationships through pricing and service quality. Working paper, Massachusetts Institute of Technology, Cambridge.

Chen, H., M. Frank. 2004. Monopoly pricing when customers queue. *IIE Trans.* **36**(6) 569–581.

Clay, K. B., D. S. Sibley, P. Srinagesh. 1992. Ex post vs. ex ante pricing: Optimal calling plans and tapered tariffs. *J. Regulatory Econom.* **4** 115–138.

DeGraba, P. 1995. Buying frenzies and seller-induced excess demand. *RAND J. Econom.* **26**(2) 331–342.

De Vany, A. 1976. Uncertainty, waiting time, and capacity utilization: A stochastic theory of product quality. *J. Political Econom.* **84**(3) 523–542.

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

Essegaier, S., S. Gupta, Z. J. Zhang. 2002. Pricing access services. *Marketing Sci.* **21**(2) 139–159.

Giridharan, P. S., H. Mendelson. 1994. Free-access policy for internal network. *Inform. Systems Res.* **5**(1) 1–21.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston.

Littlechild, S. C. 1974. Optimal arrival rate in a simple queueing system. *Internat. J. Production Res.* **12**(3) 391–397.

Masuda, Y., S. Whang. 2006. On the optimality of fixed-up-to tariff for telecommunications service. *Inform. Systems Res.* **17**(3) 247–253.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Comm. ACM* **28**(3) 312–321.

Miravete, E. J. 1996. Screening consumers through alternative pricing mechanisms. *J. Regulatory Econom.* **9** 111–132.

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

Randhawa, R., S. Kumar. 2008. Usage restriction and subscription services: Operational benefits with rational customers. *Manufacturing Service Oper. Management* **10**(3) 429–447.

Scotchmer, S. 1985. Profit-maximizing clubs. *J. Public Econom.* **27**(1) 25–45.

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