M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 10, No. 3, Summer 2008, pp. 429–447
ISSN 1523-4614 | EISSN 1526-5498 | 08 | 1003 | 0429

informs®
DOI 10.1287/msom.1070.0180

Usage Restriction and Subscription Services: Operational Benefits with Rational Users

Ramandeep S. Randhawa
McCombs School of Business, University of Texas at Austin, Austin, Texas 78712
rsr@mail.utexas.edu

Sunil Kumar
Graduate School of Business, Stanford University, Stanford, California 94305
skumar@stanford.edu

This paper studies a rental firm that offers reusable products to price- and quality-of-service–sensitive customers—Netflix or Blockbuster can be thought of as the canonical example. Customers' perception of quality is determined by their likelihood of obtaining the product or service immediately upon request. We study the alternatives of offering either a subscription option that limits the number of concurrent rentals in return for a flat fee per-unit time, or a pay-per-use option with no such restriction. Customers are assumed to desire a nominal usage rate of the product, which they meet by adjusting their request rate in either option. Thus, they have a higher request rate in the subscription option. We propose a Markov chain model for customer behavior under the subscription option equivalent to the standard Poisson model under the pay-per-use option. In a large market setting, assuming exponential demand, we show that using the subscription option is more profitable for the firm. Further, via a numerical study, we show that this assumption is not essential for the result to hold. However, we show that the subscription option does not necessarily dominate the pay-per-use option in quality of service. The firm manages the trade-off between price and quality of service better in the subscription option. Moreover, we show that social welfare and the consumer surplus can also be higher in the subscription option, indicating that both the firm and the consumers can benefit from the subscription option.

Key words: subscription services; pay-per-use services; operational benefits; pricing; capacity sizing; finite customer population; loss system; on-off model; diffusion approximation
History: Received: December 19, 2005; accepted: March 23, 2007. Published online in *Articles in Advance* January 4, 2008.

1. Introduction

Many firms provide reusable products or services to customers who request these products or services in some random fashion. Moreover, the duration of product use by the customer is also random in such systems. A canonical example of such a system is a rental firm. The firm cannot always accommodate all customer requests because the number of products stocked or the number of available servers is limited. A measure of the firm's quality of service as seen by a customer is the likelihood of obtaining the product immediately upon request. To manage the variability in the presence of capacity constraints, the firm needs to set the capacity level, i.e., the number of products stocked, and the price so as to optimally extract profit.

Access to such rental firms is typically of two kinds: subscription or pay per use, with either option having an equal ease of implementation in many contexts. For example, in the case of DVD rentals, Netflix has emerged as a major player and it provides only a subscription option. On the other hand, Blockbuster mostly provides a pay-per-use option. With the introduction of Total Access, Blockbuster has also made a foray into the subscription world. Of course, one would expect a firm to do better in the face of competition by locking in customers via subscription. In this paper, we study subscription from a different angle. We study whether the firm is able to handle the inherent variability in the system better if it offers a subscription option rather than a pay-per-use option. That is, we study whether the firm receives any *operational* benefit of offering a subscription option.

We consider a rental firm serving price- and quality-sensitive customers. The firm has the option of offering a subscription or a pay-per-use option. We assume that subscribers pay a subscription fee per-unit time, independent of the usage, while the pay-per-use customers pay a price each time they use the product. In addition, we assume that subscribers are restricted to renting only one product at a time, while pay-per-use customers are not subject to such a restriction. This assumption is quite natural in a DVD rental setting, where customers could be renting multiple DVDs at a given time. The number of concurrent rentals is either contractually restricted as in the case of Netflix or not restricted as in the pay-per-use model of Blockbuster. Of course, rational customers will consider the concurrent usage restriction and react accordingly. In our paper, in both the subscription and pay-per-use settings, the firm statically sets price and capacity levels and observes demand in response. The firm's objective is to maximize the difference between the revenues it obtains from the customers and the cost of maintaining the capacity. We assume that the customer demand—that is, the number of subscribers joining the system or the rate at which pay-per-use customers show up at the firm—depends on both the prices the firm sets as well as the equilibrium likelihood of obtaining the product immediately upon request. The reader can think of this equilibrium as having been reached via repeated interactions with the customers, and, hence, this likelihood is common knowledge in the market. We solve the firm's optimization problem for both the subscription option as well as the pay-per-use option and then compare the two using an asymptotic approach. The natural asymptotic regime considered is one where the market for potential customers is large.

In comparing the two options, the first question that one might ask is: Does restricting usage in the subscription option lead to a reduction in the inherent system variability, resulting in higher quality of service at any choice of capacity or price? We show that this is not necessarily true. In fact, neither option need dominate the other in terms of quality of service. However, if profits are compared in a large market with rational price- and quality-of-service–sensitive customers, the subscription option does dominate the equivalent pay-per-use option. For the case of exponential demand, we prove that the subscription option is indeed better for the firm. Through a numerical study, we show that this result is robust with respect to the choice of demand function. Although the subscription option does not necessarily dominate the pay-per-use option on quality of service, the firm is able to better manage the trade-off between price and quality of service in the subscription option. Moreover, we show that the social welfare and the consumer surplus can also be higher in the subscription option, indicating that both the firm and the consumers can benefit from the subscription option. However, this need not always be the case, with the firm sometimes profiting from the subscription option at the customers' expense.

Our approach for solving the firm's optimization problem for either option is as follows. Denoting the size of the market—that is, the number of potential customers—by n, we first solve a nominal optimization problem at the $O(n)$ level. (As usual, $O(n)$ denotes any quantity such that $O(n)/n$ is bounded as $n \to \infty$.) In such a system, the natural scale at which stochastic variability manifests itself is $O(\sqrt{n})$. Any refinement to the nominal problem that considers the stochastic variability in the system must be at the $O(\sqrt{n})$ level. The following example may help illustrate this scale difference. Consider a newsvendor model with demand arising from n independent, identically distributed sources. In this case, the optimal stock level comprises of the mean of the total demand, which is $O(n)$, and a safety stock against variability, which is some number of standard deviations of the total demand and is $O(\sqrt{n})$. The safety stock in the newsvendor model is exactly analogous to the refinement to the nominal solution that we propose. By applying a law of large numbers argument, Proposition 1 shows that on the $O(n)$ scale, the system operates in a deterministic regime with essentially all requests being satisfied immediately. Furthermore, using a functional central limit theorem proved in earlier work, we perform a refined analysis on the $O(\sqrt{n})$ scale to obtain prices and capacity levels optimal on this scale. That is, these choices result in profits within $o(\sqrt{n})$ of the optimal values (see Proposition 4). (As usual, $o(n)$ denotes any quantity such that $o(n)/n \to 0$ as $n \to \infty$.) Furthermore, Proposition 3 shows that in this regime, pricing and capacity sizing are equivalent levers for the firm. That is, any refinement in prices can be imitated by an equivalent refinement in capacity levels for both options. This asymptotic approach allows us to compare the profits obtained under each option up to a resolution of $o(\sqrt{n})$, and we prove the primary result of the paper; that is, the profit is higher for the subscription option by a quantity that is $O(\sqrt{n})$. Combining this with the size of the possible error in estimating the profit allows us to conclude that in a large enough market, the subscription option dominates.

This paper is organized in the following manner. In §2, we propose a generic model of customer behavior when restricted to only one concurrent rental. We develop the analytical methods required to calculate the asymptotically optimal price and capacity levels that maximize the firm's profit. We solve both the nominal problem on the $O(n)$ scale and its variability refinement and show that pricing and capacity sizing are equivalent levers for the firm on the $O(\sqrt{n})$ scale. In §3, we build the framework necessary to have a unified model of unrestricted customer behavior so that we can compare the subscription and pay-per-use options. We show that this unified model reduces the model developed in §2 under usage restriction, and is equivalent to a standard Poisson model in a pay-per-use setting without such restrictions. We solve the firm's optimization problem in pay-per-use settings using the machinery developed in §2. Section 4 compares the subscription option with an equivalent pay-per-use option and proves the main result of the paper; that is, the firm's profits are higher when offering a subscription option. Section 5 discusses the conclusions and scope for future work. Finally, the appendix contains a summary of the asymptotic results in Randhawa (2006) used in this paper and the proofs of all the results.

1.1. Literature Review

We use the approach in Mendelson and Whang (1990) to build a micro-economic framework around our service system and study optimal pricing and capacity sizing at the system equilibrium. We characterize the system equilibrium as in Armony and Maglaras (2004) and Whitt (2003). These papers deal with multiserver queueing systems with congestion-sensitive demand but without economic considerations, and study the system equilibrium behavior using asymptotic methods. Perhaps the most related to our work is a recent paper (Maglaras and Zeevi 2003) that studies pricing and capacity selection in a multiserver queueing system with Poisson arrivals. This paper uses the Halfin-Whitt (see Halfin and Whitt 1981) asymptotic results to develop estimates for congestion levels to compute optimal price levels. Another application of optimal pricing and capacity sizing in an asymptotic regime is in Plambeck and Ward (2005), where the authors study static pricing, capacity selection, and dynamic scheduling using the traditional heavy traffic assumptions in an assemble-to-order setting. Optimal capacity sizing under fixed, exogenous demand is studied in Borst et al. (2004), where the authors use asymptotic methods to compute the optimal staffing level in a call center with the trade-off being between the agents' cost and the quality of service provided.

Although this paper deals with static pricing, work in Paschalidis and Tsitsiklis (2000) and Gallego and van Ryzin (1994) suggests that this is not a major limitation. Paschalidis and Tsitsiklis (2000) study revenue management in the context of Internet service provision using a Markov decision process framework where customers are only assumed to be price sensitive. Although they focus on dynamic pricing, an important conclusion of their study is that static pricing rules can achieve near-optimal performance. Gallego and van Ryzin (1994) derive a similar insight in the "classical" context of selling a set of goods within a finite-time horizon.

The theory of "club goods" in economic literature also compares subscription and pay-per-use settings (see Buchanan 1965, Cornes and Sandler 1986, Scotchmer 1985, Barro and Romer 1987). The treatment there is macroscopic in the sense that the *operational* aspects due to the dynamics of customer requests and rentals are not considered. The most similar paper in this literature to our work is Barro and Romer (1987), which demonstrates that in a ski-lift pricing setting, a firm can achieve the same profit by pricing per ride or charging a flat daily fee. We establish a similar result in our setting where we observe that as long as a change is not induced in customer behavior (via a usage restriction, for example), usage-based pricing and pricing independent of usage generate identical profits for the firm (see §4.5 for more details).

2. Subscription

2.1. A Model of Subscriber Behavior Under Usage Restriction

We begin by describing a natural Markovian model for customer behavior in a subscription setting where the customer is restricted to renting only one product at a time. The model is parametrized by three parameters, λ, μ, and ν. We associate each subscriber with a Markov chain having three states: on, off, and hold. The transition rate matrix is given by

$$\begin{array}{cccc} & \text{on} & \text{off} & \text{hold} \\ M = & \multicolumn{3}{c}{\left(\begin{array}{ccc} -\mu & \mu & 0 \\ \lambda 1_{\{\text{product available}\}} & -\lambda & \lambda 1_{\{\text{product not available}\}} \\ \nu 1_{\{\text{product available}\}} & 0 & -\nu 1_{\{\text{product available}\}} \end{array}\right)} \end{array}.$$

Subscribers spend an exponentially distributed amount of time with mean $1/\lambda$ in the off state, after which they request a product. If no products are available they transit to the hold state, otherwise, a product is assigned to them and they transit to the on state. In the hold state, they retry to obtain a product after every exponentially distributed amount of time with mean $1/\nu$ until a product is available. Once a product is assigned to them, they transit to the on state. In the on state, they use the product for an exponentially distributed amount of time with mean $1/\mu$, after which they return the product and transit to the off state. We assume that all the times spent by a subscriber in each state are independent, identically distributed, and independent of times in other states. We also assume that a product cannot be assigned to more than one subscriber at any time. This subscriber model is related to the classical Engset model (see Kleinrock 1975) used in telecommunications. In fact, for the special case in which the hold and off states are indistinguishable (i.e., when the rates at which a subscriber tries to obtain a product from the hold and off states are equal), the subscribers can be characterized by a two-state Markov chain; this is identical to the Engset model.

For the most part, we will assume that $\nu = \lambda$, which renders the hold and off states indistinguishable. When the assumption that $\nu = \lambda$ is not satisfied, we use the asymptotic limits derived in Randhawa and Kumar (2007) to obtain insights based on comparative statics on the retrial rate using numerical experiments. These are discussed in §5. The parameter choice λ is in fact intrinsic to the subscriber in the sense that each subscriber chooses λ to meet a nominal usage level. We elaborate on this intrinsic model in §3, where we shall see the importance of the underlying model in comparing different contracts the firm offers its subscribers.

We now build a demand function for the subscribers. We consider a market with a total of *n potential subscribers*. In accordance with Naor's original idea (see Naor 1969), we assume that each potential subscriber has a value S for each service completed. This value is assumed to be the same for every service completed for each potential subscriber but may differ among potential subscribers. The valuations are distributed according to a distribution $F(\cdot)$ and with density $f(\cdot)$. We shall assume that the density is strictly positive on $(0, \infty)$ and there is a value p^e such that $dx\bar{F}(x)/dx < 0$ for $x \geq p^e$, where $\bar{F}$ denotes the tail of the cumulative density function. This assumption is equivalent to the demand function corresponding to F being elastic beyond the price p^e and ensures a nontrivial solution to the pricing and capacity sizing problem. We assume that subscribers are individually rational, i.e., they join the system only if joining ensures them a positive benefit.

Subscribers' sensitivity to the quality of service is captured through the denial probability; that is, the steady-state probability that upon request a subscriber does not receive the product. We define the denial probability γ by

$$\gamma = \lim_{t\to\infty} \frac{\text{No. of denied attempts by time } t}{\text{No. of attempts by time } t}.$$

The higher the denial probability the longer it takes a subscriber to obtain the product. In particular, denote the mean time between successful attempts to obtain a product in steady-state for a subscriber by τ. We can then characterize τ as a function of the denial probability as follows.

LEMMA 1. *The mean time between successful attempts to obtain a product in steady-state for a subscriber is* $\tau = 1/\lambda + \gamma/(\nu(1-\gamma)) + 1/\mu$.

All the proofs, including that of this result, can be found in §A.2.

We assume that a potential subscriber joins the system only if the value obtained per service, S, exceeds average cost per service, as follows. When the system manager sets a subscription fee of p per-unit time, he or she only joins the system if $S > p\tau$. Therefore, the number of potential subscribers who join the system at price p per-unit time and denial probability γ, $N(p, \gamma)$, can be modeled as

$$N(p,\gamma) = n\mathbb{P}\left(S > \left(\frac{1}{\lambda} + \frac{\gamma}{\nu(1-\gamma)} + \frac{1}{\mu}\right)p\right)$$
$$= n\bar{F}\left(\left(\frac{1}{\lambda} + \frac{\gamma}{\nu(1-\gamma)} + \frac{1}{\mu}\right)p\right).$$

In building this model, we have assumed that the subscriber valuations S among the potential subscribers are not random draws but are precisely distributed according to the distribution F. That is, as is usually done with demand curves, we assume that for any $x > 0$, the fraction of the subscribers who have valuations that exceed x is exactly $\bar{F}(x)$.

In addition to the subscription fee, the system manager also decides on the capacity level k. Assuming each product costs \$$c$ per-unit time, where $c > 0$, the optimization problem of the system manager can be stated as

$$\max_{(p,\,k)\in\mathbb{R}_+\times\mathbb{Z}_+} \quad pN(p,\gamma) - kc$$
$$\text{s.t.} \quad \gamma = d(N(p,\gamma), k),$$

where $d(N(p,\gamma), k)$ is the denial probability as a function of the number of subscribers and the capacity level. Note that for any subscription fee and capacity level set, the denial probability must satisfy an equilibrium condition, which is captured via the constraint $\gamma = d(N(p,\gamma), k)$. The inability to characterize this equilibrium denial probability in a simple form renders this problem, as stated, extremely hard to solve. This motivates us to try to approximately solve it in the natural asymptotic regime of large number of potential subscribers. In particular, we shall let n grow without bound and use the superscript n to make explicit the dependence of the various parameters on n, i.e., p^n, k^n, and γ^n denote the subscription fee, number of products, and denial probability, respectively, and $N^n(p^n, \gamma^n)$ denotes the number of potential subscribers joining the system. The optimization problem can then be written as

$$\max_{(p^n,\,k^n)\in\mathbb{R}_+\times\mathbb{Z}_+} \quad \Pi^n(p^n, k^n) \equiv p^n N^n(p^n, \gamma^n) - k^n c$$
$$\text{s.t.} \quad \gamma^n = d(N^n(p^n,\gamma^n), k^n).$$

With this setup, we try to find a sequence of subscription fees and number of products (p^{n*}, k^{n*}) optimal in the limiting regime as $n \to \infty$. We shall begin by providing a nominal solution, and then constructing a refinement that accounts for the variability. The nominal solution optimizes the objective function on the $O(n)$ scale by appealing to the law of large numbers, while use of the central limit theorem implies that the refinement is on the $O(\sqrt{n})$ scale. The $O(\sqrt{n})$ scale refinement can be understood by considering a newsvendor model with demand arising from n independent, identically distributed sources. In this case, the optimal stock level comprises of the mean of the total demand, which is $O(n)$, and a safety stock against variability, which is some number of standard deviations of the total demand and is $O(\sqrt{n})$.

A sequence (p^{n*}, k^{n*}) is said to be a nominal solution to the problem if for all sequences (p^n, k^n)

$$\liminf_{n\to\infty} \frac{\Pi^n(p^{n*}, k^{n*})}{n} \geq \limsup_{n\to\infty} \frac{\Pi^n(p^n, k^n)}{n}.$$

We denote the corresponding nominal profit by $\overline{\Pi}$, i.e., $\overline{\Pi} = \liminf_{n\to\infty} \Pi^n(p^{n*}, k^{n*})/n$. The nominal solution is a deterministic solution, and thus can be refined by studying the variability in the system. To do so, we introduce the notion of asymptotic optimality at the $O(\sqrt{n})$ scale. We expect the actual profit in the system to be less than that expected in the nominal setting. This motivates us to look at the loss in profits due to variability and refine our solution to minimize these losses. That is, for any sequence (p^n, k^n), denoting $\widetilde{\Pi}^n(p^n, k^n) \equiv \sqrt{n}(\overline{\Pi} - \Pi^n(p^n, k^n)/n)$ as the $O(\sqrt{n})$ scale loss in profit, which is the difference between the nominal profit $\overline{\Pi}$ and that obtained while using the rule (p^n, k^n), we define an asymptotically optimal sequence as follows.

Definition 1. A sequence (p^{n*}, k^{n*}) is said to be asymptotically optimal if the scaled loss in profit is minimized, that is, for any sequence (p^n, k^n)

$$\limsup_{n\to\infty} \widetilde{\Pi}^n(p^{n*}, k^{n*}) \leq \liminf_{n\to\infty} \widetilde{\Pi}^n(p^n, k^n).$$

It is clear that for a sequence to be asymptotically optimal, it must be a nominal solution.

2.2. Nominal Solution

Henceforth, we shall assume that $\nu = \lambda$, with the general retrial case discussed in §5. We first loosely illustrate how the nominal solution can be characterized by the solution to a static, deterministic, optimization problem. Then, we prove a precise statement characterizing nominal solutions in Proposition 1. Scaling the objective function by n and letting n grow without bound should result in a deterministic problem, and, hence, one expects that its solution should correspond to one where $\gamma = 0$.[1] Further, for $\gamma = 0$, the capacity level must equal the offered load in the system, which is defined as the mean number of products that will be in use in steady-state when the system has infinite capacity. We expect the probability that a subscriber will be using a product in steady-state is $\lambda/(\lambda+\mu)$. Hence, as the number of subscribers who join this system is $n\bar{F}(p^n/m)$, where $m \equiv \lambda\mu/(\lambda+\mu)$, the offered load is $n(\lambda/(\lambda+\mu))\bar{F}(p^n/m)$. This implies that if the scaled price and capacity levels converge as $(p^n, k^n/n) \to (p, k)$, then the capacity level k must equal $(\lambda/(\lambda+\mu))\bar{F}(p/m)$. Thus, we obtain the following static optimization problem, which we call the nominal problem.

$$\begin{aligned} \max_{(p,k)\in\mathbb{R}_+^2} \quad & p\bar{F}\left(\frac{p}{m}\right) - kc \\ \text{s.t.} \quad & k = \frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{p}{m}\right). \end{aligned} \tag{1}$$

The assumptions on F ensure an interior solution. Thus, we use the necessary condition for optimality, that is, the first derivative of the objective function must be zero, to establish that the optimal subscription fee $\bar{p}$ satisfies

$$\bar{F}\left(\frac{\bar{p}}{m}\right) = \frac{1}{m} f\left(\frac{\bar{p}}{m}\right)\left(\bar{p} - \frac{\lambda}{\lambda+\mu}c\right), \tag{2}$$

and the optimal capacity level is given by $\bar{k} = (\lambda/(\lambda+\mu))\bar{F}(\bar{p}/m)$. The following result makes the loose reasoning above precise.

Proposition 1. *For any sequence (p^n, k^n) to be a nominal solution, $\gamma^n \to 0$, $p^n \to \bar{p}$, and $k^n/n \to \bar{k}$ as $n \to \infty$.*

[1] Before scaling, γ is $O(1/\sqrt{n})$, and hence tends to zero as n grows without bound. We shall refine the nominal solution using the right order of γ in §2.3.

2.3. Refining the Nominal Solution

We begin by defining the capacity imbalance in the system as the difference between the number of products and the offered load scaled by n. As derived earlier, this offered load is $n(\lambda/(\lambda+\mu))\bar{F}(p^n/m)$. Thus, for any sequence (p^n, k^n), the capacity imbalance is given by

$$\theta^n = \frac{k^n}{n} - \frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{p^n}{m}\right). \tag{3}$$

The following lemma presents a "decay" condition on the asymptotically optimal imbalance.

Lemma 2. *For any asymptotically optimal sequence (p^n, k^n), $\limsup_{n\to\infty} |\theta^n\sqrt{n}| < \infty$.*

We now focus on constructing an asymptotically optimal sequence (p^n, k^n) with the associated equilibrium denial probability γ^n. Motivated by the above lemma, we restrict our search to θ^n such that $\lim_{n\to\infty}\theta^n\sqrt{n} = \theta$, where $\theta \in \mathbb{R}$. We choose $p^n = \bar{p} + \phi^n$ and $k^n = (\bar{k} + \kappa^n)n$ such that (p^n, k^n, θ^n) satisfy (3) and $\phi^n, \kappa^n \to 0$. In fact, we shall assume without loss of generality that $\phi^n\sqrt{n} \to \phi$ and $\kappa^n\sqrt{n} \to \kappa$. We can rewrite (3) using the Taylor series expansion of $\bar{F}(p^n/m)$ about $(\bar{p}/m)$ to obtain

$$\theta^n = \frac{\phi^n}{\mu} f\left(\frac{\bar{p}}{m}\right) + \kappa^n + o(\phi^n), \tag{4}$$

and hence

$$\theta = \frac{\phi}{\mu} f\left(\frac{\bar{p}}{m}\right) + \kappa.$$

We are now ready to asymptotically characterize the denial probability for our system.

Proposition 2. *For a sequence (p^n, k^n) such that $\theta \equiv \lim_{n\to\infty}\theta^n\sqrt{n}$ exists, $\gamma^n\sqrt{n} \to \gamma$ given by*

$$\gamma = \sqrt{\frac{\lambda+\mu}{m\bar{F}(\bar{p}/m)}}\, h\left(-\left[\theta + f\left(\frac{\bar{p}}{m}\right)\bar{p}\frac{\gamma}{\lambda+\mu}\right]\sqrt{\frac{\lambda+\mu}{m\bar{F}(\bar{p}/m)}}\right), \tag{5}$$

where h is the hazard rate function associated with the standard normal distribution.

Note that in the limit the denial probability solves a fixed-point relation and is a function of only the limiting capacity imbalance. In the following result, we show that the profit is also a function of the capacity imbalance alone, and, hence, pricing and capacity sizing are equivalent levers for the firm at the $O(\sqrt{n})$ scale.

Proposition 3. *For a sequence (p^n, k^n) such that $p^n = \bar{p} + \phi^n$, $k^n = (\bar{k} + \kappa^n)n$ and $\theta \equiv \lim_{n\to\infty} \theta^n \sqrt{n}$ exists, the loss in profit function satisfies*

$$\lim_{n\to\infty} \widetilde{\Pi}^n(p^n, k^n) = \theta c + \bar{p}^2 f\left(\frac{\bar{p}}{m}\right)\frac{\gamma}{\lambda}.$$

Therefore, any sequence (p^n, k^n) that has the same limiting imbalance θ has the same asymptotic loss in profit.

This result implies that the asymptotically optimal sequence is characterized by the value of θ that minimizes $\lim_{n\to\infty} \widetilde{\Pi}^n(p^n, k^n)$. Hence, the optimal asymptotic capacity imbalance θ^* solves

$$\min_{\theta\in\mathbb{R}} \quad \theta c + \bar{p}^2 f\left(\frac{\bar{p}}{m}\right)\gamma/\lambda$$

$$\text{s.t.} \quad \gamma = \sqrt{\frac{\lambda+\mu}{m\bar{F}(\bar{p}/m)}}\, h\left(-\left[\theta + f\left(\frac{\bar{p}}{m}\right)\bar{p}\frac{\gamma}{\lambda+\mu}\right] \cdot \sqrt{\frac{\lambda+\mu}{m\bar{F}(\bar{p}/m)}}\right).$$

Denoting $z \equiv h'^{-1}(c\lambda m\bar{F}(\bar{p}/m)/(\bar{p}^2(\lambda+\mu)f(\bar{p}/m) - \bar{p}f(\bar{p}/m)c\lambda))$, the first-order conditions imply that the optimal denial probability and capacity imbalance are

$$\gamma^* = \sqrt{\frac{\lambda+\mu}{m\bar{F}(\bar{p}/m)}}\, h(z)$$
$$\theta^* = -z\sqrt{\frac{m\bar{F}(\bar{p}/m)}{\lambda+\mu}} - f\left(\frac{\bar{p}}{m}\right)\bar{p}\frac{\gamma^*}{\lambda+\mu}. \tag{6}$$

Checking first-order conditions establishes optimality because h is a convex function (cf. Zeltyn and Mandelbaum 2005, p. 12).

We now characterize all asymptotically optimal sequences in the following result.

Proposition 4. *The sequence $(\bar{p} + \phi^{n*}, (\bar{k} + \kappa^{n*})n)$, where $(\phi^{n*}/\mu)f(\bar{p}/m) + \kappa^{n*} = \theta^*/\sqrt{n} + o(1/\sqrt{n})$, is asymptotically optimal.*

The equivalence of corrections in price and capacity levels suggests that the firm could choose to keep one parameter at the nominal level and correct the other. This is particularly useful for a DVD rental firm because one expects making changes to capacity levels to be easier than changing prices. Hence, in this scenario, the firm can set prices at the nominal level and make corrections in the capacity level in accordance with the equilibrium to achieve the asymptotically optimal solution. However, if changing prices is easier than changing capacity, the reverse can be done.

Corollary 1. *The sequences $(\bar{p}, (\bar{k} + \theta^*/\sqrt{n})n)$ and $(\bar{p} + (\theta^*/\sqrt{n})(\mu/f(\bar{p}/m)), \bar{k}n)$ are asymptotically optimal.*

This subscriber model is a natural model of customer behavior in a subscription setting under a concurrent contractually enforced usage restriction. When confronted with a different contract without such a restriction, one expects the rational users to modify their behavior. For example, when operating in a pay-per-use setting, one would expect the customers to reduce their frequency of requests because they can have more than one concurrent rental. This change in behavior must be factored in before we can set ourselves up to compare subscription and pay-per-use options. This is the subject of the next section.

3. Pay Per Use

3.1. A Model of Customer Behavior

Consider the DVD rental world again. Here, Netflix offers solely subscription-based service, while Blockbuster primarily uses pay-per-use pricing. In turn, Blockbuster's customers may request additional products, even though they might already be renting a product, and thus can rent many products concurrently, while Netflix imposes a restriction on the customer usage, i.e., customers are only allowed to rent a certain number of DVDs at a time. In this setting, it seems reasonable to assume that customers have an intrinsic "need" or "desire" to watch a certain number of DVDs per time unit, and that they would modify their request rates to achieve this rate depending on the option they sign up for. In this section, our goal is to build a common customer model as the basis of understanding how customer behavior changes with different options. We will then analyze the firm's decision of offering either a usage-restricting subscription option or a pay-per-use option.

We begin by defining our market. We shall assume that there are n potential customers, each characterized by

1. A benefit derived per usage/rental, S, that is random and distributed according to F.

2. A usage time for the rental of each product, which we assume to be exponentially distributed with mean $1/\mu$.

3. A nominal "desire" of renting x products per time unit. For example, each customer wants to rent x products per month. We assume that this desire is constant across customers.

4. A random attempt/request process that allows the customer to realize the desired rental level described above. We shall assume the interarrival times of the requests are exponentially distributed. Any request denied is considered lost, i.e., there are no retrials.

Before analyzing the pay-per-use option in detail, we establish that this model is consistent with the model proposed in §2.1 for the subscription option when usage is restricted.

3.2. Subscription Option: Reduction to Subscriber Model of §2

We study the subscription option that restricts a subscriber to renting only one product at a time. Any request generated while the subscriber is still renting the product is *de facto* denied, by the terms of the contract. To compensate for this, rational users will adjust their request rate. Let $\lambda(x)$ be the rate at which each subscriber generates requests in this setting. This rate depends on the nominal desire x described; we shall elaborate on this shortly.

A straightforward application of the Markovian property of the exponential distribution leads to the following result.

Lemma 3. *Subscribers generating requests according to a renewal process with exponential interarrival times with mean* $1/\lambda(x)$, *where requests generated while a product is being rented are denied, can equivalently be modeled according to the on-off model in* §2 *of the paper.*

Hence, the subscribers defined here are exactly those in §2 of the paper, and we can directly use the analysis performed in that section. Note that as each subscriber wishes to maintain a constant usage level, he or she chooses a request rate $\lambda(x)$ so that the nominal usage (under no capacity constraints) equals x, i.e.,

$$\frac{\lambda(x)\mu}{\lambda(x)+\mu} = x. \tag{7}$$

For a given x, we can compute the corresponding rate of turning On, λ, and then perform the analysis in §2 to compute the optimal prices the firm must set and the corresponding number of products.

We shall now analyze the pay-per-use option.

3.3. Pay-Per-Use Option

Under the pay-per-use option, the firm allows the customer to rent unlimited multiple products, i.e., a customer request is always accepted subject to available capacity. The customer pays per usage in return. Let $\beta(x)$ denote the rate at which each customer generates requests in this setting. Further, because customers wish to maintain a constant usage level, they choose a request rate $\beta(x)$ so that the nominal usage (under no capacity constraints) equals x, i.e.,

$$\beta(x) = x. \tag{8}$$

We can mimic the arguments in §2 to solve the firm's optimization problem when offering a pay-per-use option. Readers may choose to skip this section and jump directly to the comparison of the two options in §4.

For convenience, we shall drop the dependence of $\beta(x)$ on x. We now build a demand model for the pay-per-use customers. As in §2, let γ denote the steady-state denial probability. When the system manager sets a pay-per-use price of p_p per use, each customer joins the system only if the expected value obtained by joining the system, which is the probability of obtaining the product times the value derived from using the product, $(1-\gamma)(S-p_p)$ is positive, i.e., $S > p_p$. Thus, at a price per usage p_p, $N_p(p_p) = n\mathbb{P}(S > p_p)$ customers join the option. Note that the number of pay-per-use customers joining the system does not depend on the denial probability. This is because we have imposed no costs on customers for checking the availability of the products; i.e., if it was costly for customers to arrive at the firm, there would be an implicit cost associated with the denial and this would affect the demand. Although the demand is independent of the quality of service, the actual requests accepted still depend on it, and so do the firm's profits. Noting that customers pay for the products per usage, it is more convenient to talk of the total arrival rate of customers, as opposed to the number of customers joining the system. Note that the customers' requests arrive according to a Poisson process with the total arrival rate Λ given as a function of the pay-per-use price by

$$\begin{aligned}\Lambda(p_p) &= \beta N_p(p_p)\\ &= n\beta \bar{F}(p_p).\end{aligned}$$

The optimization problem of the system manager can be stated as

$$\max_{(p_p,\,k_p)\in\mathbb{R}_+^2} \quad p_p(1-\gamma)\Lambda(p_p) - k_p c$$

$$\text{s.t.} \quad \gamma = d(\Lambda(p_p), k_p),$$

where $d(\Lambda(p_p), k_p)$ is the denial probability as a function of the arrival rate of the customers and the number of products.

As before, we shall use the superscript n to make explicit the dependence of the various parameters on n, i.e., p_p^n, k_p^n, γ^n, and $\Lambda(p_p^n)$ denote the pay-per-use fee, the capacity level, the denial probability, and the rate of arrival of the pay-per-use customers, respectively. The optimization problem can then be written as

$$\max_{(p_p^n,\,k_p^n)\in\mathbb{R}_+^2} \quad \Pi^n(p_p^n, k_p^n) \equiv p_p^n(1-\gamma^n)\Lambda^n(p_p^n) - k_p^n c$$

$$\text{s.t.} \quad \gamma^n = d(\Lambda^n(p_p^n), k_p^n).$$

With this setup, we try to find a sequence of prices and number of products (p_p^{n*}, k_p^{n*}) optimal in the limiting regime as $n \to \infty$.

3.3.1. Nominal Solution. We begin by solving the nominal problem for this system analogous to (1) given by

$$\begin{aligned}&\max_{(p_p,\,k_p)\in\mathbb{R}_+^2} \quad p_p\beta\bar{F}(p_p) - k_p c\\ &\text{s.t.} \quad k_p = \frac{\beta}{\mu}\bar{F}(p_p).\end{aligned} \tag{9}$$

It is easy to verify using the first-order conditions that the nominal price $\bar{p}_p$ satisfies the relation

$$\bar{F}(\bar{p}_p) = \left(\bar{p}_p - \frac{c}{\mu}\right) f(\bar{p}_p) \tag{10}$$

and the nominal capacity level $\bar{k}_p = (\beta/\mu)\bar{F}(\bar{p}_p)$. We then have the following result.

Proposition 5. *A sequence (p_p^n, k_p^n) is a nominal solution if $\gamma^n \to 0$ and $(p_p^n, k_p^n/n) \to (\bar{p}_p, \bar{k}_p)$ as $n \to \infty$.*

3.3.2. Refining the Nominal Solution. As before, we define the capacity imbalance in the system as the difference between the number of products and the offered load in the absence of denials scaled by n. Thus, for any sequence (p_p^n, k_p^n), the capacity imbalance is given by

$$\theta^n = \frac{k_p^n}{n} - \frac{\beta}{\mu}\bar{F}(p_p^n). \tag{11}$$

We now focus on constructing an asymptotically optimal sequence (p_p^n, k_p^n) with the associated denial probability γ^n. An analog to Lemma 2 holds in this setting as well, which motivates us to restrict our search to θ^n such that $\lim_{n\to\infty}\theta^n\sqrt{n} = \theta$, where $\theta \in \mathbb{R}$. We choose $p_p^n = \bar{p}_p + \phi_p^n$ and $k_p^n = (\bar{k}_p + \kappa_p^n)n$ such that (p_p^n, k_p^n, θ^n) satisfy (11) and $\phi_p^n, \kappa_p^n \to 0$. Further, we rewrite (11) using the Taylor series expansion of $\bar{F}(p_p^n)$ about $\bar{p}_p$ to obtain

$$\theta^n = \frac{\phi_p^n\beta}{\mu} f(\bar{p}_p) + \kappa_p^n + o(\phi_p^n).$$

The asymptotic denial probability can now be characterized as follows.

Proposition 6. *For a sequence (p_p^n, k_p^n) such that $\theta \equiv \lim_{n\to\infty}\theta^n\sqrt{n}$ exists, $\gamma^n\sqrt{n} \to \gamma$ given by*

$$\gamma = \sqrt{\frac{\mu}{\beta\bar{F}(\bar{p}_p)}}\, h\left(-\theta\sqrt{\frac{\mu}{\beta\bar{F}(\bar{p}_p)}}\right). \tag{12}$$

We now show that as in the case of the subscription option, profit is a function of the capacity imbalance alone, and hence pricing and capacity sizing are equivalent levers for the firm at the $O(\sqrt{n})$ scale.

Proposition 7. *For a sequence (p_p^n, k_p^n) such that $p_p^n = \bar{p}_p + \phi_p^n$, $k_p^n = (\bar{k} + \kappa_p^n)n$ and $\theta \equiv \lim_{n\to\infty}\theta^n\sqrt{n}$ exists, the loss in profit function satisfies*

$$\lim_{n\to\infty} \widetilde{\Pi}^n(p_p^n, k_p^n) = \theta c + \bar{F}(\bar{p}_p)\beta\bar{p}_p\gamma.$$

This result implies that the asymptotically optimal sequence is characterized by the value of θ that minimizes $\lim_{n\to\infty}\widetilde{\Pi}^n(p_p^n, k_p^n)$. Hence, the optimal asymptotic capacity imbalance θ^* solves

$$\min_{\theta\in\mathbb{R}} \quad \theta c + \bar{F}(\bar{p}_p)\beta\bar{p}_p\gamma$$

$$\text{s.t.} \quad \gamma = \sqrt{\frac{\mu}{\beta\bar{F}(\bar{p}_p)}}\, h\left(-\theta\sqrt{\frac{\mu}{\beta\bar{F}(\bar{p}_p)}}\right).$$

As before, we use the first-order conditions to characterize the optimal solution. Hence, denoting $z \equiv h'^{-1}(c\bar{F}(\bar{p}_p)/(\bar{F}(\bar{p}_p)\bar{p}_p\mu))$, the optimal denial probability and capacity imbalance are given by

$$\gamma^* = \sqrt{\frac{\mu}{\beta\bar{F}(\bar{p}_p)}}\,h(z)$$
$$\theta^* = -z\sqrt{\frac{\beta\bar{F}(\bar{p}_p)}{\mu}}. \tag{13}$$

We now characterize all asymptotically optimal sequences in the following result.

Proposition 8. *The sequence $(\bar{p}_p + \phi_p^{n*}, (\bar{k}_p + \kappa_p^{n*})n)$, where $(\phi_p^{n*}\beta/\mu)f(\bar{p}_p) + \kappa_p^n = \theta^*/\sqrt{n} + o(1/\sqrt{n})$, is asymptotically optimal.*

Having solved the firm's optimization problem for both the subscription and pay-per-use options, we now turn our attention to their comparison. We will also compare measures like social welfare and consumer surplus in these options.

4. Comparing the Two Options

We shall compare the subscription and pay-per-use options in the asymptotic regime of large markets. Recall that customers choose their request rates in each scenario to achieve their nominal usage level. Thus, comparing (7) and (8), for a meaningful comparison of the two settings, we must have

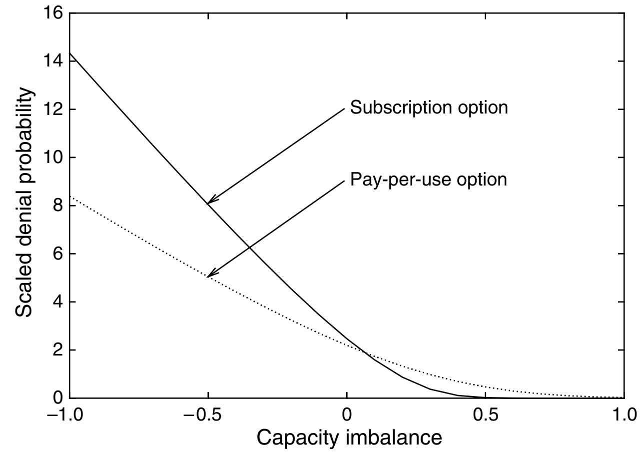
$$\frac{\lambda\mu}{\lambda+\mu} = \beta. \tag{14}$$

We begin by comparing the nominal profits in the two options. Because the two settings are statistically equivalent at the nominal level, we expect no difference in their nominal profits. Indeed, comparing (1–2) and Proposition 1 with (9–10) and Proposition 5, we obtain the following result.

Proposition 9. *The nominal profit in the subscription option equals that in the pay-per-use option.*

Now, we investigate how each option handles the variability in the system manifested at the $O(\sqrt{n})$ level. The question of each option's ability to handle variability is moot if one option always has better quality of service at *every* load. We begin by showing that one option does not dominate the other on quality of service. Given this, we shall compare the profits in the two options. We shall then perform a finer analysis to ascertain which option results in higher profits. This analysis consists of comparing the asymptotic loss in profit (at the $O(\sqrt{n})$ level) in both options. Finally, we shall compare the consumer surplus and social welfare in these options.

Figure 1 Comparison of Quality of Service in the Two Options



4.1. Quality of Service: Neither Option Dominates

We compute the asymptotic quality-of-service levels (using Propositions 2 and 6) for both options at different capacity imbalance levels and compare them in Figure 1. We set $F(x) = 1 - e^{-x}$ for $x \geq 0$, $\lambda = 5$, and $\mu = \frac{5}{4}$ (so that $m = 1$) for these computations. The figure illustrates that neither option dominates the other. At low capacity imbalance levels, the pay-per-use option offers a higher quality of service, while the subscription option is better at higher imbalance levels. Figure 1 can be explained loosely as follows. As the number of subscribers who are on increases, the number of subscribers attempting to obtain products decreases. Consequently, at high capacity levels, the denial probability seen by the subscribers is smaller than that seen by the pay-per-use customer stream whose attempt rate is independent of the state of the system. Similarly, the denial probability can be explained to be higher in the subscription option at low capacity levels.

4.2. Firm's Profits: Subscription Option Always Dominates

We now focus on establishing that the firm's profits are higher in the subscription option. We will normalize $m = 1$ for convenience. We shall compare the asymptotic loss in profit in the two options. Noting that h'^{-1} is a function that does not have a closed form, proving such a claim in complete generality is quite difficult. Instead, we prove this result for a specific demand function and will provide numerical results for a few others. Denoting the optimal capacity imbalances in the subscription and pay-per-use options by θ_s^* and θ_p^*, respectively, we have the following result.

Proposition 10. *For the exponential demand function with $F(x) = e^{-\alpha x}$ for $x > 0$, for any constant $\alpha > 0$, the firm's asymptotic loss in profit is higher in the pay-per-use option. Further, the capacity imbalance is lower in the subscription option, $\theta^{s*} < \theta^{p*}$.*

4.2.1. Comparison of Firm's Profits: Numerical Experiments. We now provide numerical results demonstrating that the profits are higher when the firm offers a subscription option for two demand functions: a linear demand function and a Pareto demand function with constant price elasticity.

1. *Linear demand*: To obtain a linear demand function, we choose a uniform distribution for the valuations, i.e., $F(x) = x$ for $0 \le x \le 1$. We numerically compute the loss in profit for a large number of parameter values. In particular, we vary both λ and the cost per product c, while choosing μ so that $m = 1$. In each instance, we observed that the loss in profit as well as the capacity imbalance were lower in the subscription option. As a specific example, Table 1 compares the optimal imbalances and the loss in profits for the two options as the cost per product varies with $\lambda = 2$. We observe that the subscription option has a lower capacity imbalance and realizes a lower loss in profit.

2. *Iso-elastic demand*: We choose a Pareto distribution for the valuations, i.e., $F(x) = 1 - x^{-2}$ for $x \ge 1$, which has a price elasticity of 2. Table 2 compares the optimal imbalances and the loss in profits for the two options as λ varies with the cost per product set at $c = 1$. We again choose μ so that $m = 1$. We observe that the subscription option has a lower capacity imbalance and realizes a lower loss in profit.

Table 1 Subscription vs. Pay Per Use: Linear Demand

	Subscription		Pay per use	
c	θ_s^*	Loss in profit	θ_p^*	Loss in profit
0.01	1.070	0.012	1.516	0.017
0.5	0.109	0.263	0.389	0.371
1.0	−0.704	0.298	−0.051	0.422
1.5	−2.766	0.202	−0.396	0.286

Table 2 Subscription vs. Pay Per Use: Pareto Demand

	Subscription		Pay per use	
λ	θ_s^*	Loss in profit	θ_p^*	Loss in profit
1.1	−0.658	2.400	0.913	2.517
2	−0.208	0.759	0.389	1.073
5	−0.104	0.379	0.308	0.848
10	−0.069	0.253	0.290	0.800

4.3. Comparison of Consumer Surplus and Social Welfare

We shall construct examples for the exponential demand function to show that the consumer surplus and social welfare, which is the sum of the consumer surplus and the firm's profit, can be higher or lower in the subscription option as compared to the pay-per-use option.

We shall begin by computing the consumer surplus in the two options. We shall normalize $m = 1$ as before and denote $\bar{p} \equiv \bar{p}_s = \bar{p}_p$, where $\bar{p}_s$ and $\bar{p}_p$ denote the nominal prices in the subscription and pay-per-use options, respectively. For a system with n subscribers, the consumer surplus can be written out as

$$CS_s^n = n \int_{(p^n(1+\gamma^n/\lambda))}^{\infty} \bar{F}(p)\, dp$$
$$= \frac{e^{-\alpha\bar{p}}}{\alpha} n - \left(\sqrt{n}\phi_s^{n*} + \sqrt{n}\frac{\gamma^n \bar{p}}{\lambda}\bar{p}\right) e^{-\alpha\bar{p}}\sqrt{n}.$$

Similarly, we can write out the consumer surplus for the pay-per-use customers as

$$CS_p^n = n(1-\gamma^n) \int_{p^n}^{\infty} \bar{F}(p)\, dp$$
$$= \frac{e^{-\alpha\bar{p}}}{\alpha} n - \left(\sqrt{n}\phi_p^{n*} + \sqrt{n}\gamma^n\right) e^{-\alpha\bar{p}}\sqrt{n}.$$

Hence,

$$\lim_{n\to\infty} \frac{CS_s^n}{n} = \lim_{n\to\infty} \frac{CS_p^n}{n} = \frac{e^{-\alpha\bar{p}}}{\alpha} \equiv \overline{CS}.$$

Because the firm's profit can be completely characterized by the value of the capacity imbalance θ, using Propositions 4 and 8, the firm can choose to

Table 3 **Comparison of Consumer Surplus and Social Welfare**

	Subscription		Pay per use	
c	Loss in CS	Loss in SW	Loss in CS	Loss in SW
0.1	0.012	0.097	0.016	0.136
1	0.172	0.611	0.163	0.783
5	0.647	1.123	0.261	0.935
10	0.530	0.735	0.125	0.415

Table 4 **Request Rates Used in the Experiments**

	Maximum products rented out				
Experiment no.	1	2	3	5	∞ (Pay per use)
1	0.4	0.34	0.334	0.333	0.333
2	2	1.098	1.013	1	1
3	10	2.12	1.75	1.67	1.67

set any value of ϕ_i^{n*}, $i = s, p$, and adjust the value of κ^{n*} to obtain the same profit. Hence, denoting the loss in surplus (due to variability) from the nominal value as $\widehat{CS}_i = \lim_{n\to\infty} \sqrt{n}(\overline{CS} - CS_i^n/n)$, we see that the firm can decrease $\widehat{CS}_i$ in an unbounded fashion by increasing ϕ_i^* arbitrarily. Hence, the customers obtain a higher consumer surplus when the firm sets lower prices. This result can be explained in a different manner. Note that if the firm sets any particular price, it can adjust the capacity level so that the capacity imbalance, and hence the quality of service, remains unchanged. Thus, reducing prices while maintaining the same quality of service increases the consumer surplus.

To compare the consumer surplus and social welfare in the two options, we shall fix the price levels at the nominal values, i.e., set $\phi_i^{n*} = 0$ for $i = s, p$. Table 3 displays the losses in consumer surplus and social welfare for different values of the parameter c at fixed $\lambda = 2$, prices set at the nominal values, and $\alpha = 1$. We observe that for low values of c, the loss in these measures are lower in the subscription option, while this is reversed for high values of c.

4.4. Multiple but Limited Concurrent Rentals

Thus far, we have considered the two extreme cases of allowing just one or unlimited concurrent rentals. In reality, Netflix allows subscribers to rent between one and four DVDs concurrently. So we consider what happens if the firm allows subscribers to hold a limited number of concurrent rentals, in which the limit is larger than one.

If r concurrent rentals are allowed among n customers, we need to keep track of an r state Markov chain, and asymptotically, we expect to get an r dimensional diffusion process. This is sufficiently complicated to make analysis difficult: We choose to perform a numerical study to quantify the effect of easing the usage restriction in this manner. As r increases, subscribers would decrease the rate at which they generate requests to meet their nominal desire x. So their nominal usage rate and the consequent nominal profits realized by the firm will not change. Comparing asymptotic profits reduces to comparing asymptotic loss in profit. Using simulations, we compute the optimal loss in profit for the cases where the subscribers are allowed to rent up to five products concurrently. The simulations are performed for a system with $n = 1{,}000$ potential subscribers, with the mean rental duration of $1/\mu = 0.5$. We chose λ to be 2, 0.4 and 10 for $r = 1$, and computed λ for the other values of r accordingly. Table 4 lists the values of the request rates used in the different settings to ensure the same nominal usage rate.

We use an exponential demand function, i.e., $F(x) = 1 - e^{-x}$, $x \geq 0$ and use simulations to estimate the profit for 50 different price values and then pick the optimal profit. Appealing to the equivalence of pricing and capacity sizing, we stick with the nominal capacity level. For each price, one needs to estimate the denial probability in an iterative manner because the number of subscribers depends on the denial probability and vice versa. We do this by beginning with a denial probability of 0 and then performing a run of 10,000 time units to estimate the denial probability in this system. We repeat this run 20 times to estimate γ with an observed maximum ratio of standard deviation to the mean of 1%. We then update our estimate of the denial probability and repeat the simulation until there is a difference of 0 or 1 in the number of people joining the system under the old and new denial probability values. The results of this study are illustrated in Table 5. Each number displayed in the table corresponds to the loss in profit from the nominal amount (given by the solution to (1)) scaled by $\sqrt{n}$. We observe that the scaled loss in profit is increasing as the number of concurrent rentals allowed increases.

Table 5 Loss in Profit When Multiple Products Can Be Rented at a Time

	Maximum products rented out				
Experiment no.	1	2	3	5	∞ (Pay per use)
1	0.32	0.34	0.37	0.37	0.37
2	0.44	0.58	0.63	0.64	0.64
3	0.31	0.68	0.79	0.83	0.82

Thus, the simulations seem to suggest that restricting customer usage in the subscription option allows the firm to generate higher profits, and the more restrictive the usage in the subscription option, the higher the profits.

4.5. Importance of Usage Restriction

We highlight that the operational benefit of subscription lies in the manner in which the customer behavior is restricted by the contract, rather than in the pricing mechanism implied by the contract. To demonstrate this, we shall look at a market with n potential customers with a random valuation S per service. Each customer's behavior is governed by an underlying Markov chain as in §2 under any pricing contract. Let us first consider the contract wherein the customers are charged per-unit time; we shall refer to this as contract A (this is the setting in §2). Let $\xi(\cdot)$ denote the rate at which services are completed as a function of the quality of service γ; $\xi(\gamma) = 1/\tau$ with τ as defined in Lemma 1. If a fee p_A per-unit time is charged, then a potential subscriber will join the system if the benefit rate exceeds the cost rate, i.e., $S\xi(\gamma) > p_A$. Thus, the demand function under this contract can be written as $N_A(p_A, \gamma) = n\bar{F}(p_A/\xi(\gamma))$.

Now, let us consider the per-usage pricing contract—call it contract B. Again, potential users sign up for this contract if the benefit rate exceeds the cost rate, i.e., they join if $S\xi(\gamma) > p_B\xi(\gamma)$. Thus, the demand function under this contract is given by $N_B(p_B, \gamma) = n\bar{F}(p_B)$.

The following result proves that contracts A and B generate the same amount of profits for the firm.

Lemma 4. *For any market size n, contracts A and B are revenue equivalent. That is, any profit generated by the firm using one contract can be replicated using the other contract, albeit at a different price.*

5. Discussion

This paper provides a technique for computing prices and capacity levels that approximately maximize a large rental firm's profit. This technique is used to characterize the asymptotically optimal solution in subscription and pay-per-use environments, and it is shown that the firm's profit is higher in the subscription setting.

This paper considers a special case of the subscriber's retrial rate, i.e., $\nu = \lambda$. The general case is quite difficult to analyze, even asymptotically (see §3 of Randhawa and Kumar 2007). We shall use numerical experiments to perform comparative statics on the retrial rate. In particular, as Table 6 demonstrates, for a linear demand function with a cost per product $c = 1$, increasing the retrial rate leads to an increase in the firm's profits. Although increasing the retrial rate increases the denial probability, it also enables the subscribers to obtain service quicker, and, hence, increasing the value they obtain by joining the option. It is this latter effect that dominates and allows the firm to extract higher profits. An analytical analysis of this phenomenon is a topic for future work.

As the retrial rate increases, we expect the system to begin to behave more like a queueing system, where, upon being denied service, people wait in a queue. Using simulations, we estimate the ratio of the denial rate to the retrial rate and compare it to the expected queue length in a queueing system. Table 7 displays these estimates along with their 95% confidence intervals. Observe that this ratio is quite close to the expected queue length estimate. This makes a case for approximating the economic model in a queueing system via a loss model or vice versa.

In this paper we focused on the benefits of a subscription option as opposed to a pay-per-use option. However, a firm might wish to offer both options in parallel. This case can be handled in a straightforward manner using the tools developed in this paper.

Table 6 Firm's Asymptotic Loss in Profit with 95% Confidence Intervals for General Retrial Rates

	Retrial rate, ν				
θ	0.1	1	2	3	4
−0.5	0.548 ± 0.014	0.329 ± 0.012	0.289 ± 0.008	0.289 ± 0.008	0.287 ± 0.013
0	0.594 ± 0.014	0.388 ± 0.010	0.340 ± 0.006	0.323 ± 0.008	0.318 ± 0.004
0.5	0.686 ± 0.011	0.566 ± 0.003	0.544 ± 0.002	0.534 ± 0.002	0.529 ± 0.002

Table 7 Using Increasing Retrial Rates to Model a Queue

n	$\nu = 10$	$\nu = 100$	$\nu = 1000$	Queue
10	8.033 ± 0.094	8.010 ± 0.0622	7.998 ± 0.06	8.000 ± 0.015
100	20.838 ± 0.551	20.065 ± 0.345	20.092 ± 0.338	20.0213 ± 0.129
1,000	67.068 ± 1.092	64.236 ± 1.078	63.922 ± 1.99	63.980 ± 0.455

In fact, we can show that the equivalence of pricing and capacity sizing carries through to this setting as well.

As a concluding note, we comment on the distributional assumptions made in this paper. Although we have made Markovian assumptions, we contend that all the results for the subscription option hold even for general distributions of on and off times, as long as the distributions have a density. In fact, these results depend on just the means of these distributions. Cohen (1957) (and Randhawa and Kumar 2007) proves that the invariant distribution for the number of products in use depends only on the first moment of the on and off times. Regarding the pay-per-use option, we cannot do away with the Poisson arrival process assumption, but the service times can have an underlying general distribution. This can be proved by using the dependence of the invariant distribution of an $M/G/k/k$ system on just the first moment of the service distribution (see Gross and Harris 1998, pp. 245–247).

Appendix

Before embarking on the proof of the results in this paper, we shall state some asymptotic results, proved in Randhawa (2006), that will be useful.

A.1. Asymptotic Results

Consider a sequence of systems with the nth system consisting of $\lceil an + b^n\sqrt{n}\rceil > 0$ subscribers, where $a > 0$ and $b^n \in \mathbb{R}$ such that $b^n \to b$. Let $Q^n(t)$ represent the number of products in use at time t in the system when there are n subscribers, i.e., $Q^n(t) = \sum_{i=1}^{\lceil an+b^n\sqrt{n}\rceil} 1_{\{\text{subscriber } i \text{ is on at time } t\}}$. We consider capacity levels of the form $k^n = \bar{k}n + \kappa^n\sqrt{n}$ for some $\bar{k} \in \mathbb{R}_+$ and $\kappa^n \in \mathbb{R}$ such that $\kappa^n \to \kappa$. Define $q^n(\cdot) = Q^n(\cdot)/n$, the centered and scaled process

$$\hat{Q}^n(t) = \frac{Q^n(t) - k^n}{\sqrt{n}} \le 0$$

and $\bar{q}(\cdot) = \min(\bar{k}, (\lambda/(\lambda + \mu)a)$. We then have the following asymptotic results for this system (Lemma 20 in Appendix C of Randhawa 2006).

Lemma 5. *If $q^n(0) \to \bar{q}(0)$ a.s., then*

1. $q^n \to \bar{q}$.
2. *If $\bar{k} = (\lambda/(\lambda + \mu))a$ and $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$, then $\hat{Q}^n \Rightarrow \hat{Q}$, where $\hat{Q}(\cdot)$ is a reflected affine-drift diffusion process with an upper reflecting barrier at* 0. *That is,*

$$\hat{Q}(t) = \hat{Q}(0) - (\lambda+\mu)\int_0^t \left(\hat{Q}(s) + \kappa - \frac{\lambda}{\lambda+\mu}b\right) ds + \sqrt{2am}B(t) - Y(t),$$

where B is a standard Brownian motion and Y is the nonnegative, nondecreasing process such that $\int_0^t \hat{Q}(u)\, dY(u) = 0$ and $\hat{Q}(t) \le 0$, $\forall\, t \ge 0$ and $Y(0) = 0$.

3. *The invariant distribution of $\hat{Q}^n(\cdot)$, $\hat{\pi}^n \to \hat{\pi}$, where $\hat{\pi}$ is the unique invariant distribution of the diffusion process $\hat{Q}$. Further, $\mathbb{E}_{\hat{\pi}^n}\hat{Q}^n \to \mathbb{E}_{\hat{\pi}}\hat{Q}$.*
4. *The density corresponding to $\hat{\pi}$ is given by*

$$\hat{p}(x) = \exp\left(-\frac{1}{2am}(\lambda+\mu)\left(x+\kappa-\frac{\lambda}{\lambda+\mu}b\right)^2\right) \cdot \left[\int_{-\infty}^0 \exp\left(-\frac{1}{2am}(\lambda+\mu)\left(z+\kappa-\frac{\lambda}{\lambda+\mu}b\right)^2\right)dz\right]^{-1}, \quad x \le 0.$$

Suppose instead of the subscribers, there is an exogenous stream of customers who arrive according to a Poisson process with rate $\lambda_p = \lambda_1 n + \lambda_2^n\sqrt{n}$, where $\lambda_1 > 0$ and $\lambda_2^n \in \mathbb{R}$ such that $\lambda_2^n \to \lambda_2$. Let $Q(\cdot) \in D_{\mathbb{R}}[0, \infty)$ be the process that denotes the number of products in use by the exogenous customers. Define $q^n(\cdot)$ and $\hat{Q}^n$ as before and $\bar{q}(\cdot) \equiv \min(\bar{k}, \lambda_1/\mu)$. We then have the following asymptotic results for this system (Lemma 21 in Appendix C of Randhawa 2006).

Lemma 6. *If $q^n(0) \to \bar{q}(0)$ a.s., then*

1. $q^n \to \bar{q}$.
2. *If $\bar{k} = \lambda_1/\mu$ and $\hat{Q}^n(0) \Rightarrow \hat{Q}(0)$, then $\hat{Q}^n \Rightarrow \hat{Q}$, where $\hat{Q}(\cdot)$ is a reflected affine-drift diffusion process with an upper reflecting barrier at* 0. *That is,*

$$\hat{Q}(t) = \hat{Q}(0) + \lambda_2 t - \mu\int_0^t (\hat{Q}(s) + \kappa)\, ds + \sqrt{2\lambda_1}B(t) - Y(t),$$

where B is a standard Brownian motion and Y is the nonnegative, nondecreasing process such that $\int_0^t \hat{Q}(u)\, dY(u) = 0$ and $\hat{Q}(t) \le 0$, $\forall\, t \ge 0$ and $Y(0) = 0$.

3. *The invariant distribution of $\hat{Q}^n(\cdot)$, $\hat{\pi}^n \to \hat{\pi}$, where $\hat{\pi}$ is the unique invariant distribution of the diffusion process $\hat{Q}$. Further, $\mathbb{E}_{\hat{\pi}^n}\hat{Q}^n \to \mathbb{E}_{\hat{\pi}}\hat{Q}$.*
4. *The density corresponding to $\hat{\pi}$ is given by*

$$\hat{p}(x) = \exp\left(-\frac{1}{2}\frac{\mu}{\lambda_1}\left(x+\kappa-\frac{\lambda_2}{\mu}\right)^2\right) \cdot \left[\int_{-\infty}^0 \exp\left(-\frac{1}{2}\frac{\mu}{\lambda_1}\left(z+\kappa-\frac{\lambda_2}{\mu}\right)^2\right)dz\right]^{-1}, \quad x \le 0.$$

Armed with Lemmas 5 and 6, we are now ready to prove the results of this paper.

A.2. Proofs

A.2.1. Proofs of Results in §2. We shall begin by deriving an expression for the denial probability for the general retrial case that will be quite useful in establishing Lemma 1. Let $Q_1(t)$ and $Q_2(t)$ denote the number of subscribers in the on and hold states, respectively, and π be the invariant distribution associated with the Markov process $Q(\cdot) = (Q_1(\cdot), Q_2(\cdot))$. Further, we shall use the notation $\mathbb{E}_\pi Q_i \equiv \mathbb{E}[Q_i(t) \mid Q_i(0) = X]$ for $i = 1, 2$ and $t \geq 0$, where X is a random variable distributed according to π.

Lemma 7. *For a system with n subscribers with associated steady-state loss probability γ, we have*

$$\gamma = \frac{\lambda n - (\lambda + \mu)\mathbb{E}_\pi Q_1 + (\nu - \lambda)\mathbb{E}_\pi Q_2}{\lambda(n - \mathbb{E}_\pi Q_1) + (\nu - \lambda)\mathbb{E}_\pi Q_2}.$$

Proof. We can write $Q_1(t)$, for any $t > 0$, as follows

$$Q_1(t) = Q_1(0) + A\left(\lambda \int_0^t (n - Q_1(s) - Q_2(s))\, ds\right) + R\left(\nu \int_0^t Q_2(s)\, ds\right) - D\left(\mu \int_0^t Q_1(s)\, ds\right) - Y(t), \quad \text{a.s.}, \quad (15)$$

where $A(\cdot)$, $R(\cdot)$, and $D(\cdot)$ are three independent Poisson processes with unit rate, corresponding to the attempts by subscribers in the off state, retrials by subscribers in the hold state and service completions by subscribers in the on state, and

$$\begin{aligned} Y(\cdot) = {} & A\left(\lambda \int_0^{\cdot} (n - Q_1(s) - Q_2(s))\, ds\right) + R\left(\nu \int_0^{\cdot} Q_2(s)\, ds\right) \\ & - A\left(\lambda \int_0^{\cdot} 1_{\{Q_1(s)<k\}} (n - Q_1(s) - Q_2(s))\, ds\right) \\ & - R\left(\nu \int_0^{\cdot} 1_{\{Q_1(s)<k\}} Q_2(s)\, ds\right) \end{aligned}$$

counts the number of denied attempts. Using renewal theory arguments (see Proposition 3.3.1 in Ross 1996), we obtain

$$\begin{aligned} & \lim_{t\to\infty} \frac{A\left(\lambda \int_0^t (n - Q_1(s) - Q_2(s))\, ds\right)}{t} \\ & \quad = \lambda n - \lambda \lim_{t\to\infty} \frac{1}{t}\int_0^t Q_1(s)\, ds - \lambda \lim_{t\to\infty} \frac{1}{t}\int_0^t Q_2(s)\, ds \\ & \quad = \lambda(n - \mathbb{E}_\pi Q_1 - \mathbb{E}_\pi Q_2). \end{aligned} \quad (16)$$

Similarly, we get

$$\lim_{t\to\infty} \frac{R\left(\nu \int_0^t Q_2(s)\, ds\right)}{t} = \nu \mathbb{E}_\pi Q_2 \quad (17)$$

$$\lim_{t\to\infty} \frac{D\left(\mu \int_0^t Q_1(s)\, ds\right)}{t} = \mu \mathbb{E}_\pi Q_1. \quad (18)$$

Using (16–18) in (15), we obtain

$$\lim_{t\to\infty} \frac{Y(t)}{t} = \lambda n - (\lambda + \mu)\mathbb{E}_\pi Q_1 + (\nu - \lambda)\mathbb{E}_\pi Q_2.$$

Thus, we obtain

$$\begin{aligned} \gamma &= \lim_{t\to\infty} \frac{Y(t)}{A\left(\lambda \int_0^t (n - Q_1(s) - Q_2(s))\, ds\right) + R\left(\nu \int_0^t Q_2(s)\, ds\right)} \\ &= \frac{\lambda n - (\lambda + \mu)\mathbb{E}_\pi Q_1 + (\nu - \lambda)\mathbb{E}_\pi Q_2}{\lambda(n - \mathbb{E}_\pi Q_1) + (\nu - \lambda)\mathbb{E}_\pi Q_2}, \end{aligned}$$

which proves the claim. □

This result immediately implies the following.

Corollary 2. *For a system with n subscribers and retrial rate $\nu = \lambda$, the associated loss probability γ is given by*

$$\gamma = \frac{\lambda n - (\lambda + \mu)\mathbb{E}_\pi Q}{\lambda(n - \mathbb{E}_\pi Q)},$$

where $Q(t)$ denotes the number of subscribers in on state at time t and π is the invariant distribution of $Q(\cdot)$.

Proof of Lemma 1. The mean time between successful attempts to obtain a product in steady state for a subscriber can be written as

$$\begin{aligned} & \lim_{t\to\infty} \frac{t}{\text{No. of successful attempts per subscriber by } t} \\ & \quad = \lim_{t\to\infty} \frac{nt}{A\left(\int_0^t \mu Q_1(s)\, ds\right)} \\ & \quad = \frac{n}{\mu \mathbb{E}_\pi Q_1}. \end{aligned} \quad (19)$$

We shall now obtain a relation between $\mathbb{E}_\pi Q_1$ and $\mathbb{E}_\pi Q_2$, which will allow us to compute the mean time between successful attempts as a function of γ, λ, ν, and μ.

For the case $k > 0$, we shall first compute the limiting fraction of time spent by any subscriber in the off state relative to the on state, i.e.,

$$\lim_{t\to\infty} \frac{\text{Time spent by subscriber } i \text{ in off state by } t}{\text{Time spent by subscriber } i \text{ in on state by } t}.$$

Let $T_{off}(t)$ and $T_{on}(t)$ denote the times spent by subscriber i in the off and on states, respectively. Let $\{\tau_{on}^j, \tau_{off}^j\}$ be the sequence of times spent in the on and off states, respectively, by this subscriber; i.e. $\{\tau_{on}^j\}$ and $\{\tau_{off}^j\}$ are sequences of independent, exponentially distributed random variables with means $1/\mu$ and $1/\lambda$, respectively. Let $S(t)$ denote the number of services completed by subscriber i by time t. Then, we have the following

$$\lim_{t\to\infty} \frac{\sum_{j=1}^{S(t)} \tau_{off}^j}{\sum_{j=1}^{S(t)} \tau_{on}^j + \tau_{on}^{S(t)+1}} \leq \lim_{t\to\infty} \frac{T_{off}(t)}{T_{on}(t)} \leq \lim_{t\to\infty} \frac{\sum_{j=1}^{S(t)} \tau_{off}^j + \tau_{off}^{S(t)+1}}{\sum_{j=1}^{S(t)} \tau_{on}^j}.$$

As each visit to the hold state is for a finite amount of time, $S(t)\uparrow\infty$ as $t\to\infty$. Thus, we obtain

$$\lim_{t\to\infty}\frac{T_{off}(t)}{T_{on}(t)}=\frac{\mu}{\lambda}.$$

Using this result, we can compute

$$\mathbb{E}_\pi Q_1=\frac{\lambda}{\lambda+\mu}(n-\mathbb{E}_\pi Q_2). \tag{20}$$

Observe that this relation holds for the case $k=0$ trivially. The result now follows by combining (20) and Lemma 7 in (19). $\square$

Henceforth, all the proofs shall be for the case $\nu=\lambda$ where the hold and off states are indistinguishable. Thus, the system will be described by $Q^n(\cdot)$, where $Q^n(t)$ is the number of subscribers in the on state at time t when there are a total of n potential subscribers. Let π^n be the invariant distribution of $Q^n(\cdot)$.

Proof of Proposition 1. Consider any sequence (p^n,k^n). Choose a subsequence n' such that $\gamma^{n'}\to\tilde{\gamma}$, $p^{n'}\to\tilde{p}$, and $k^{n'}/n'\to\tilde{k}$ as $n'\to\infty$. For this scenario,

$$\lim_{n'\to\infty}\frac{\Pi^{n'}(p^{n'},k^{n'})}{n'}=\tilde{p}\bar{F}\left(\left(\frac{1}{\lambda(1-\tilde{\gamma})}+\frac{1}{\mu}\right)\tilde{p}\right)-\tilde{k}c,$$

where use of Corollary 2 implies

$$\tilde{\gamma}=\frac{\lambda\bar{F}\left(\left(\frac{1}{\lambda(1-\tilde{\gamma})}+\frac{1}{\mu}\right)\tilde{p}\right)-(\lambda+\mu)\tilde{k}}{\lambda\left(\bar{F}\left(\left(\frac{1}{\lambda(1-\tilde{\gamma})}+\frac{1}{\mu}\right)\tilde{p}\right)-\tilde{k}\right)}.$$

Consider the following optimization problem

$$\max_{(p,k)\in\mathbb{R}_+^2}\ p\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right)p\right)-kc$$
$$\text{s.t.}\quad \gamma=\frac{\lambda\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right)p\right)-(\lambda+\mu)k}{\lambda\left(\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right)p\right)-k\right)}. \tag{21}$$

We will show that the above problem is maximized at $(\bar{p},\bar{k})$ and has $\gamma=0$, which proves the claim.

Fix the number of products at k. Reorganizing terms in the expression for γ in (21), we obtain

$$\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right)p\right)=k\mu\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right).$$

The optimization problem can thus be rewritten as

$$\max_{\gamma\ge 0}k\mu\bar{F}^{-1}\left(k\mu\left(\frac{1}{\lambda(1-\gamma)}+\frac{1}{\mu}\right)\right)-kc.$$

It is easy to see that the above is maximized at $\gamma=0$, and hence (21) has the same solution as (1), which is $(\bar{p},\bar{k})$. $\square$

We adapt the proof technique in Plambeck and Ward (2005) to prove the following result.

Proof of Lemma 2. We shall prove that for any sequence (p^n,k^n) with the associated γ^n if $\limsup_{n\to\infty}|\sqrt{n}\theta^n|=\infty$, then $\limsup_{n\to\infty}\widetilde{\Pi}^n(p^n,k^n)=\infty$, and hence (p^n,k^n) cannot be asymptotically optimal.

Given (p^n,k^n) we can write

$$\begin{aligned}\widetilde{\Pi}^n(p^n,k^n)&=\sqrt{n}\left(\bar{\Pi}-\left(p^n\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)}+\frac{1}{\mu}\right)p^n\right)-\frac{k^n}{n}c\right)\right)\\&=\sqrt{n}\left(\bar{\Pi}-\left(p^n\bar{F}\left(\frac{p^n}{m}\right)-p^n\left(\bar{F}\left(\frac{p^n}{m}\right)\right.\right.\right.\\&\qquad\left.\left.\left.-\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)}+\frac{1}{\mu}\right)p^n\right)\right)\right)\right)+\frac{k^n}{\sqrt{n}}c\\&\overset{(a)}{=}\sqrt{n}\ \bar{\Pi}-\sqrt{n}\left(p^n\bar{F}\left(\frac{p^n}{m}\right)-\frac{k^n}{n}c\right)\\&\quad+\sqrt{n}p^n\left(\bar{F}\left(\frac{p^n}{m}\right)-\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)}+\frac{1}{\mu}\right)p^n\right)\right)\\&\ge\sqrt{n}\left(\theta^nc+p^n\left(\bar{F}\left(\frac{p^n}{m}\right)\right.\right.\\&\qquad\left.\left.-\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)}+\frac{1}{\mu}\right)p^n\right)\right)\right),\end{aligned} \tag{22}$$

where the inequality follows because the second term in (a) is upper bounded by the solution to the problem

$$\max_{(p,k)\in\mathbb{R}_+^2}\ p\bar{F}\left(\frac{p}{m}\right)-kc$$
$$\text{s.t.}\quad k=\frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{p}{m}\right)+\theta^n,$$

which is $\bar{\Pi}-\theta^nc$ at $p=\bar{p}$.

We now divide the proof based on two cases.

Case 1: $\limsup_{n\to\infty}\sqrt{n}\theta^n=\infty$. Because $\gamma^n\ge 0$, this implies that the second term in the lower bound for $\widetilde{\Pi}^n(p^n,k^n)$ in (22) is nonnegative. Hence,

$$\widetilde{\Pi}^n(p^n,k^n)\ge\sqrt{n}\theta^nc,$$

and hence $\limsup_{n\to\infty}\widetilde{\Pi}^n(p^n,k^n)=\infty$.

Case 2: $\liminf_{n\to\infty}\sqrt{n}\theta^n=-\infty$. Consider a subsequence n' such that $\lim_{n'\to\infty}\sqrt{n'}\theta^{n'}=-\infty$.

Using (3), we split each capacity imbalance $\theta^{n'}$ into corrections in prices and number of products, i.e., set $p^{n'}=\bar{p}+\phi^{n'}$, and $k^{n'}=(\bar{k}+\kappa^{n'})n'$, such that

$$\theta^{n'}=\frac{\phi^{n'}}{\mu}f\left(\frac{\bar{p}}{m}\right)+\kappa^{n'}+o(\phi^{n'}).$$

Using Corollary 2, Taylor's expansion for $N^{n'}(p^{n'},\gamma^{n'})/n'=\bar{F}((1/(\lambda(1-\gamma^{n'}))+1/\mu)p^{n'})$ about $\bar{p}/m$, and because $Q^{n'}(\cdot)\le(\bar{k}+\kappa^{n'})n'$, we have the following relation for n' large enough

$$\begin{aligned}\gamma^{n'}\ge\lambda n'\left(\bar{F}\left(\frac{\bar{p}}{m}\right)-f\left(\frac{\bar{p}}{m}\right)\frac{\phi^{n'}}{m}-f\left(\frac{\bar{p}}{m}\right)\bar{p}\gamma^{n'}/\lambda\right.\\\left.+o(\phi^{n'})+o(\gamma^{n'})\right)\cdot[\lambda(N^{n'}(p^{n'},\gamma^{n'})-\mathbb{E}_{\pi^{n'}}Q^{n'})]^{-1}\end{aligned}$$

$$
\begin{aligned}
&-(\lambda+\mu)n'\left(\frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{\bar{p}}{m}\right)+\kappa^{n'}\right)\\
&\cdot[\lambda(N^{n'}(p^{n'},\gamma^{n'})-\mathbb{E}_{\pi^{n'}}Q^{n'})]^{-1}\\
\geq&\left[-\lambda f\left(\frac{\bar{p}}{m}\right)\frac{\phi^{n'}}{m}-f\left(\frac{\bar{p}}{m}\right)\bar{p}\gamma^{n'}+o(\phi^{n'})+o(\gamma^{n'})\right.\\
&\left.-(\lambda+\mu)\kappa^{n'}\right]\cdot\left[\left(\lambda\bar{F}\left(\frac{\bar{p}}{m}\right)-\lambda\frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{\bar{p}}{m}\right)\right)+o(1)\right]^{-1}\\
=&\left[-(\lambda+\mu)\theta^{n'}-f\left(\frac{\bar{p}}{m}\right)\bar{p}\gamma^{n'}+o(\gamma^{n'})\right]\\
&\cdot\left[m\bar{F}\left(\frac{\bar{p}}{m}\right)+o(1)\right]^{-1}, \qquad (23)
\end{aligned}
$$

where the second inequality follows as $\lim_{n'\to\infty}p^{n'}=\bar{p}$, $\lim_{n'\to\infty}k^{n'}/n'=\bar{k}$, and

$$\lim_{n'\to\infty}\mathbb{E}_{\pi^{n'}}Q^{n'}/n'=\frac{\lambda}{\lambda+\mu}\bar{F}\left(\frac{\bar{p}}{m}\right)=\bar{k},$$

and $o(1)$ represents a function that converges to 0 as n increases without bound.

We can simplify (23) to obtain

$$\frac{\theta^{n'}}{\gamma^{n'}}\geq-\frac{1}{\lambda+\mu}\left(m\bar{F}\left(\frac{\bar{p}}{m}\right)+f\left(\frac{\bar{p}}{m}\right)\bar{p}+o(\gamma^{n'})/\gamma^{n'}+o(1)\right). \qquad (24)$$

Because $\lim_{n'\to\infty}\sqrt{n'}\theta^{n'}=-\infty$, the above implies

$$\lim_{n'\to\infty}\sqrt{n'}\gamma^{n'}=\infty.$$

Using Taylor's expansion in (22), and then using (24) we get

$$
\begin{aligned}
&\lim_{n'\to\infty}\tilde{\Pi}^{n'}(p^{n'},k^{n'})\\
&\geq\lim_{n'\to\infty}\sqrt{n'}\left(\theta^{n'}c+(\bar{p}+\phi^{n'})f\left(\frac{\bar{p}}{m}\right)\bar{p}\gamma^{n'}\cdot\frac{1}{\lambda}+o(\gamma^{n'})\right)\\
&\geq\lim_{n'\to\infty}\sqrt{n'}\gamma^{n'}\left(-\frac{c}{\lambda+\mu}\left(m\bar{F}\left(\frac{\bar{p}}{m}\right)+f\left(\frac{\bar{p}}{m}\right)(\bar{p})\right)\right.\\
&\qquad\left.+\bar{p}^2f\left(\frac{\bar{p}}{m}\right)\cdot\frac{1}{\lambda}+O(\phi^{n'})+o(\gamma^{n'})/\gamma^{n'}+o(1)\right)\\
&\overset{(b)}{=}\left(\lim_{n'\to\infty}\sqrt{n'}\gamma^{n'}\frac{f(\bar{p}/m)}{\lambda}\right)\left(\bar{p}-\frac{\lambda c}{\lambda+\mu}\right)^2,
\end{aligned}
$$

where (b) follows by using (2). Again referring to (2), we must have $\bar{p}-\lambda c/(\lambda+\mu)>0$. Thus, we obtain $\lim_{n'\to\infty}\tilde{\Pi}^{n'}(p^{n'},k^{n'})=\infty$, which implies

$$\limsup_{n\to\infty}\tilde{\Pi}^{n}(p^{n},k^{n})=\infty. \quad \square$$

Proof of Proposition 2. We shall first compute the asymptotic denial probability when the number of subscribers joining the system is independent of the denial probability. Once we have this limit, we shall incorporate the dependence to complete the proof.

Consider a sequence of systems with $N^n=\lceil an+b^n\sqrt{n}\rceil\geq 0$ subscribers in the nth system, where $a>0$ and $b^n\in\mathbb{R}$ such that $b^n\to b$, and number of products given by $k^n=[\lambda/(\lambda+\mu)]an+\kappa\sqrt{n}$ for some $\kappa\in\mathbb{R}$. We can then characterize the asymptotic scaled denial probability for this system as follows.

Lemma 8. *The scaled denial probability converges as follows.*

$$\lim_{n\to\infty}\gamma^n\sqrt{n}=\sqrt{\frac{\lambda+\mu}{am}}\;h\left(-\left(\kappa-\frac{\lambda}{\lambda+\mu}b\right)\sqrt{\frac{\lambda+\mu}{am}}\right).$$

We now let the number of subscribers in the nth system be $N^n(p^n,\gamma^n)$. To complete the proof, we need to establish that $\gamma^n\sqrt{n}$ converges. To see this convergence, note that $N^n(p^n,\gamma^n)\leq N^n(p^n,0)$, and, hence, performing a birth-death chain analysis, we obtain $\gamma^n=d(N^n(p^n,\gamma^n),k^n)\leq d(N^n(p^n,0),k^n)$. A direct application of Lemma 8 allows us to conclude that $d(N^n(p^n,0),k^n)\sqrt{n}$ converges as both $\phi^n\sqrt{n}$ and $\kappa^n\sqrt{n}$ converge. Hence, $\limsup_{n\to\infty}\gamma^n\sqrt{n}<\infty$. Further, as the relation (5) must hold for any convergent subsequence of $\gamma^n\sqrt{n}$, the result follows. $\square$

Proof of Lemma 8. Arguing as in Corollary 2, we can show that

$$\gamma^n=\frac{\lambda N^n-(\lambda+\mu)\mathbb{E}_{\pi^n}Q^n}{\lambda\,(N^n-\mathbb{E}_{\pi^n}Q^n)}.$$

Hence, using Lemma 5(c), we have

$$\gamma^n\sqrt{n}\to-\frac{(\lambda+\mu)^2}{a\lambda\mu}\mathbb{E}_{\hat{\pi}}\left(\hat{Q}+\kappa-\frac{\lambda}{\lambda+\mu}b\right).$$

Using Lemma 5.3, we can compute

$$\mathbb{E}_{\hat{\pi}}\left(\hat{Q}+\kappa-\frac{\lambda}{\lambda+\mu}b\right)=\sqrt{\frac{am}{\lambda+\mu}}h\left(-\left(\kappa-\frac{\lambda}{\lambda+\mu}b\right)\sqrt{\frac{\lambda+\mu}{am}}\right).$$

Hence, we have

$$\lim_{n\to\infty}\gamma^n\sqrt{n}=\sqrt{\frac{\lambda+\mu}{am}}\;h\left(-\left(\kappa-\frac{\lambda}{\lambda+\mu}b\right)\sqrt{\frac{\lambda+\mu}{am}}\right). \quad \square$$

Proof of Proposition 3. We can express $\tilde{\Pi}^n(p^n,k^n)$ as

$$
\begin{aligned}
\tilde{\Pi}^n(p^n,k^n)&=\sqrt{n}\left(\bar{\Pi}-\left(p^n\bar{F}\left(\left(\frac{1}{\lambda(1-\gamma^n)}+\frac{1}{\mu}\right)p^n\right)-(k^n/n)c\right)\right)\\
&=\sqrt{n}\left(\theta^nc+\bar{p}^2f\left(\frac{\bar{p}}{m}\right)(\gamma^n/\lambda)+o(1/\sqrt{n})\right),
\end{aligned}
$$

where the second equality follows by application of Taylor's expansion and using (4). Using $\theta=\lim_{n\to\infty}\theta^n\sqrt{n}$ and Proposition 2, we obtain

$$\lim_{n\to\infty}\tilde{\Pi}^n(p^n,k^n)=\theta c+\bar{p}^2f\left(\frac{\bar{p}}{m}\right)\gamma/\lambda,$$

where γ satisfies (5). $\square$

A.2.2. Proofs of Results in §3.

Proof of Lemma 3. The product is always rented for an exponentially distributed amount of time with mean $1/\mu$. Thus, we only need to show that the request process is identical in the two scenarios. If the customer is not renting a product, then this equivalence trivially holds. Suppose the customer is renting the product. Then, as requests generated before the product is returned are denied immediately, the next valid request will be the first request after returning the product. However, a straightforward application of the memoryless property of the exponential distribution implies that this must again be exponentially distributed with mean $1/\lambda(x)$. Thus, the result follows. □

Proofs of Results in §3.2. The key difference between the subscription and pay-per-use options is in the computation of the denial probability. For the pay-per-use option, the following analog of Lemma 7 holds.

Lemma 9. *For a sequence* (p_p^n, k^n) *with associated loss probability* γ^n, *we have*

$$\gamma^n = \frac{\Lambda^n(p_p^n) - \mu \mathbb{E}_{\pi^n} Q^n}{\Lambda^n(p_p^n)}.$$

Proof. We can write the number of customers using the product at time t, $Q^n(t)$, for any $t > 0$, as follows

$$Q^n(t) = Q^n(0) + A(\Lambda^n(p_p^n)t) - \left(\mu \int_0^t Q^n(s)\,ds\right) - Y^n(t) \quad \text{a.s.},$$

where $A(\cdot)$ and $D(\cdot)$ are two independent Poisson processes with unit rate, and

$$Y^n(\cdot) = A(\Lambda^n(p_p^n)\cdot) - A\left(\Lambda^n(p_p^n)\int_0^{\cdot} 1_{Q^n(s)<k^n\}}\,ds\right)$$

counts the number of denied attempts. The result now follows by continuing as in the proof of Lemma 7. □

Using this result and Lemma 6, all other results in this section can be proved in a manner similar to those in §2 and are omitted for brevity.

A.2.3. Proofs of Results in §4.

Proof of Proposition 10. We shall use the qualifying subscript s for terms pertaining to the subscription option in §2. The nominal solution is given by $\bar{p}_s = \bar{p}_p = 1/\alpha + c/\mu \equiv \bar{p}$. Using this, we characterize the optimal capacity imbalances θ_i^* and denial probabilities γ_i^*, $i = s, p$ for the two options. (6) implies that

$$\gamma_s^* = \sqrt{(\lambda+\mu)e^{\alpha\bar{p}}}\; h(z),$$

where $z = -\sqrt{(\lambda+\mu)e^{\alpha\bar{p}}}(\theta_s^* + \alpha e^{-\alpha\bar{p}}\bar{p}(\gamma_s^*/(\lambda+\mu))$. Then using the fact that for a standard normal distribution $h'(x) = h^2(x) - xh(x)$, we can rewrite the optimality condition on z, $h'(z) = \alpha c/(\mu+\alpha c)$ in conjunction with (6) to obtain the following relation

$$(\gamma_s^*)^2\left(\frac{1+\alpha\bar{p}}{\lambda+\mu}\right)e^{-\alpha\bar{p}} + \theta_s^*\gamma_s^* = \frac{\alpha c}{\mu+\alpha c},$$

which implies

$$\gamma_s^* = (\lambda+\mu)e^{\alpha\bar{p}}\frac{-\theta_s^* + \sqrt{(\theta_s^*)^2 + \dfrac{4\alpha c(1+\alpha\bar{p})}{(\mu+\alpha c)(\lambda+\mu)}e^{-\alpha\bar{p}}}}{2(1+\alpha\bar{p})}. \tag{25}$$

Performing a similar analysis for the pay-per-use option, we first observe that the optimality conditions are identical, i.e., $h'(z) = \alpha c/(\mu+\alpha c)$. Hence, equating the corresponding z values for the two options, we obtain

$$-\sqrt{(\lambda+\mu)e^{\alpha\bar{p}}}\left(\theta_s^* + \alpha e^{-\alpha\bar{p}}\bar{p}\frac{\gamma_s^*}{\lambda+\mu}\right) = -\sqrt{\mu e^{\alpha\bar{p}}}\theta_p^*.$$

Using the value of γ_s^*, this can be simplified to obtain

$$\theta_p^* = \sqrt{\lambda}\left(\frac{1}{2}\left(\frac{2+\alpha\bar{p}}{1+\alpha\bar{p}}\right)\theta_s^* + \frac{\alpha\bar{p}}{2(1+\alpha\bar{p})}\cdot\sqrt{(\theta_s^*)^2 + \frac{4\alpha c(1+\alpha\bar{p})}{(\mu+\alpha c)(\lambda+\mu)}e^{-\alpha\bar{p}}}\right).$$

Note that this implies $\theta_p^* > \theta_s^*$ as $\lambda > 1$. Further, we can derive an analog of (25) to write out γ_p^* in terms of θ_p^* to obtain

$$\gamma_p^* = \frac{\mu e^{\alpha\bar{p}}}{2}\left(-\theta_p^* + \sqrt{(\theta_p^*)^2 + \frac{4\alpha c}{(\mu+\alpha c)\mu}e^{-\alpha\bar{p}}}\right),$$

which can be recast as a function of θ_s^* alone. Hence, the denial probabilities and the optimal capacity imbalance in the pay-per-use system can be written as a function of θ_s^*. This enables us to write out the difference in the loss of profits between the two options as

$$P(\theta_s^*) = c\theta_p^*(\theta_s^*) + e^{-\alpha\bar{p}}\bar{p}\gamma_p^*(\theta_s^*) - \left(\theta_s^* c + \alpha e^{-\alpha\bar{p}}\frac{\bar{p}^2}{\lambda}\gamma_s^*(\theta_s^*)\right).$$

The claim is equivalent to proving $P(\theta_s^*) > 0$. Noting that computing the value of θ_s^* requires us to deal with the hazard rate function, instead, we will prove that $P(\theta) > 0$ for any $\theta \in \mathbb{R}$. Via algebraic manipulations, we can show that $P(0) > 0$ and $P(\theta) = 0$ has no real roots. This along with the continuity of P completes the proof. □

Proof of Lemma 4. Suppose the firm sets a price p_A in contract A and stocks k products. Let γ_A be the corresponding quality of service. Then, the firm's profit rate is given by

$$\pi_A = p_A N_A(p_A, \gamma_A) - ck = n\bar{F}\left(\frac{p_A}{\xi(\gamma_A)}\right) - ck.$$

Now, consider an alternate setting where the firm offers customers contract B with a per-usage price $p_B = p_A/\xi(\gamma_A)$ per usage. Suppose the firm stocks the same number of products k. Then, in equilibrium, the number of customers who join the system, and thus the quality-of-service level, will be identical to that in contract A, and the firm's profit rate is given by

$$\pi_B = p_B\xi(\gamma_B)N_B(p_B, \gamma_B) - ck = p_B\xi(\gamma_A)N_A(p_A, \gamma_A) - ck = \pi_A.$$

The reverse argument can be constructed by setting $p_A = p_B \xi(\gamma_B)$, where γ_B is the quality of service under contract B. $\square$

References

Armony, M., C. Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Oper. Res.* **52** 271–292.

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

Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52** 17–34.

Buchanan, J. M. 1965. An economic theory of clubs. *Economica* **32** 1–14.

Cohen, J. W. 1957. The generalized Engset formulae. *Philips Telecomm. Rev.* **18**(4) 158–170.

Cornes, R., T. Sandler. 1986. *The Theory of Externalities, Public Goods, and Club Goods*. Cambridge University Press, Cambridge, UK.

Gallego, G., G. van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Sci.* **40** 999–1020.

Gross, D., C. Harris. 1998. *Fundamentals of Queueing Theory.* John Wiley & Sons, New York.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Kleinrock, L. 1975. *Queueing Systems, Vol. I: Theory*. John Wiley & Sons, New York.

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

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* **38** 870–883.

Naor, P. 1969. The regulation of queue sizes by levying tolls. *Econometrica* **37** 15–24.

Paschalidis, I., J. Tsitsiklis. 2000. Congestion-dependent pricing of network services. *IEEE/ACM Trans. Networking* **8** 171–184.

Plambeck, E. L., A. R. Ward. 2006. Optimal control of a high volume assemble-to-order system. *Math. Oper. Res.* **31** 453–477.

Randhawa, R. S. 2006. Operations with subscribers. Ph.D. thesis, Stanford University, Stanford, CA.

Randhawa, R. S., S. Kumar. 2007. Multi-server loss systems with subscribers. Working paper.

Ross, S. M. 1996. *Stochastic Processes*, 2nd ed. John Wiley & Sons, New York.

Scotchmer, S. 1985. Two-tier pricing of shared facilities in a free-entry equilibrium. *RAND J. Econom.* **16** 456–472.

Whitt, W. 2003. How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* **51** 531–542.

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *QUESTA* **51**(3/4) 361–402.