OPERATIONS RESEARCH
Vol. 57, No. 6, November–December 2009, pp. 1347–1359
ISSN 0030-364X | EISSN 1526-5463 | 09 | 5706 | 1347

DOI 10.1287/opre.1080.0629

Dynamics of New Product Introduction in Closed Rental Systems

Achal Bassamboo
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208,
a-bassamboo@northwestern.edu

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

Ramandeep S. Randhawa
McCombs School of Business, The University of Texas at Austin, Austin, Texas 78712,
ramandeep.randhawa@mccombs.utexas.edu

We study a rental system where a fixed number of heterogeneous users rent one product at a time from a collection of reusable products. The online DVD rental firm Netflix provides the motivation. We assume that rental durations of each user are independent and identically distributed with finite mean. We study transient behavior in this system following the introduction of a new product that is desired by all the users. We represent the usage process for this new product in terms of an empirical distribution. This allows us to characterize the asymptotic behavior of the usage process as the number of users increases without bound, via appropriate versions of Glivenko-Cantelli and Donsker's theorems. Analyzing the usage process, we demonstrate that an increase in the variability of the rental duration distribution can actually help the firm by allowing it to set lower capacity levels to provide a desired quality of service. Further, we show that the firm is better off not imposing any deadlines for the return of the product.

Subject classifications: queues: limit theorems, nonstationary.
Area of review: Stochastic Models.
History: Received June 2007; revision received December 2007; accepted May 2008. Published online in *Articles in Advance* March 11, 2009.

1. Introduction

This paper is motivated by the online DVD rental business: Netflix is the established incumbent, with Blockbuster making a recent foray into this space. There are several special features of this business from an operational perspective. First, the firms serve a closed population of possibly heterogeneous customers. Second, the firm limits the number of DVDs that a customer can have at any given time. In Netflix, this "max-out" number varies from one to four. Third, each customer provides a preference list. On the return of a DVD, the firm sends out the available DVD that is highest on the preference list. It is reasonable to assume that the firm has sufficient variety and quantity of "classic" movies to ensure that there is always an available DVD that can be sent out when a customer returns a previous rental. Finally, it is also reasonable to assume that no customer wants to see the same movies more than once.

The key aspect of this system is the dynamics of usage following the introduction of a new "hot" release: Analyzing steady-state behavior, assuming that all rental units are substitutes is moot. Demand for this new release is necessarily transient: Everyone watches a movie once, and eventually everyone has seen the movie. The transient process that we study is the *usage* process. For us, usage is the number of copies of the new release that are with customers at any given time, assuming that there are sufficient copies of the new release. Of course, it suffices to have as many copies as there are customers. To study usage, we need to keep track of returns of old movies by customers who have not yet seen the new release, which constitutes the request process for the new release, as well as returns of copies of the new movie. Depending on the relationship between these two components, it could very well be that peak usage is much smaller than the number of customers. Thus, analyzing the usage process allows us to decide on the number of copies to stock in order to ensure a given quality of service, as measured by stockout probabilities or fill rates. Getting insights on the behavior of the usage process, its dependence on the rental behavior of the customers, and its impact on stocking decisions, is the focus of this paper.

We model the firm as having a fixed population of n customers, all of whom desire the hot new product (movie). (In Netflix, customers are encouraged to reveal their desire to see a new release via their preference list before the actual release. As a result, it is easy to identify the population who wishes to see the movie, and ignore all others.) The customers differ with respect to their rental duration

distribution. We assume that each customer is of type θ (taking values in a finite set Θ) with probability p_θ, and a customer of type θ has a rental duration distribution F_θ. For simplicity, we assume that a customer is allowed to rent *exactly one* product at a time, and holds each such rental for a random amount of time distributed according to her F_θ. Upon returning the product, the customer *immediately* requests another product. We assume that the customer can always obtain a product that she has not previously rented. The complete model description is in §2.

We assume that the system has achieved stationarity when the new movie is introduced. At any fixed n, analyzing the usage process does not provide the structural insights we seek. Moreover, in applications like Netflix, n is typically very large. Therefore, we analyze usage in the asymptotic regime when the number of customers n increases without bound. (This asymptotic regime is akin to the multiserver limit in Halfin and Whitt 1981.) We obtain a deterministic functional law-of-large-numbers limit, the fluid limit. This limit depends not just on the first moment, but on the *entire distribution* F_θ for all $\theta \in \Theta$. This fluid limit characterizes usage on $O(n)$ scale. We provide a refinement of this limit on the $O(\sqrt{n})$ scale, obtaining a Gaussian process as a limit.

Most of the literature that analyzes multiserver limiting regimes considers either Markovian assumptions on the underlying distributions using an exponential (as in Halfin and Whitt 1981) or phase-type distribution (Puhalskii and Reiman 2000, Whitt 2005). Some exceptions to this are Krichagina and Puhalskii (1997), Armony et al. (2009), Whitt (2006), Glynn and Whitt (1991), and Reed (2007). The asymptotic limits in these systems are measure-valued processes or their functionals. Our work is in the spirit of these papers. The single-server system analyzed in processor-sharing settings in Gromoll et al. (2002) also falls into this category. Our work differs from the aforementioned papers in the method used to derive the asymptotic limits. Most of these papers utilize some form of empirical process theory for the analysis. A recent paper, Gromoll et al. (2008), uses the Glivenko-Cantelli theorem to obtain asymptotic limits for the stochastic primitives in their model, which they use to characterize fluid limits in processor-sharing queues. However, in this paper we exploit the special structure in our problem to translate the underlying queueing model into an empirical process associated with sampling n two-dimensional random variables. An application of the appropriate Glivenko-Cantelli and Donsker theorems from van der Vaart (2000) then allows us to completely characterize the limiting $O(n)$ and $O(\sqrt{n})$ processes. This is similar in spirit to Glynn and Whitt (1991), where the authors exploit the special structure in their system to derive elegant approximations with relative ease; the authors compute the asymptotic limits for an infinite server queue using the fact that the queue length for a system with deterministic service times can be easily written out as a function of the arrival counting process.

Having obtained the asymptotic behavior of usage, we turn our attention to stocking. On the dominant $O(n)$ scale, the quantity that matters for stocking is b^*, the maximum value that the fluid limit of usage attains (see §5). For the case of a homogeneous customer pool with an exponential F_θ, we show that $b^* = 1/e$. This result implies that the firm can provide excellent quality of service while stocking around 38% of the total demand. This result also illustrates the multiplexing benefit from random returns of previous rentals, and represents a result that could not have been obtained using a static analysis. We characterize the "best" distribution for a given mean that allows b^* to be made arbitrarily small; it turns out that this distribution has an arbitrarily high variance. We also study the impact of imposing return deadlines. Under the assumption that deadlines imposed are identical for all products, we show that b^* is decreasing in the deadline. Thus, it is preferable not to impose deadlines. The peak usage b^* is not monotone in the mean of the demand distribution, and so we need to use the entire dynamics of the usage process to arrive at this conclusion. In passing, we note that it is possible to envisage other implementations of deadlines where they might prove beneficial. For example, the firm may impose deadlines only for "hot" products. We do not study such settings in this paper.

When the firm wants to stock so as to achieve a desired stockout probability, the $O(\sqrt{n})$ correction comes into play. Although the limit "correction" process obtained is not very tractable, we show that only the distribution of this process at the deterministic times where the fluid limit is maximized matters. This makes the stocking decision more tractable. We provide an estimate of the stock level required to meet a given stockout probability that is accurate to a resolution of $O(\sqrt{n})$. Finally, Netflix uses a "comparable DVD" model where the user is assumed to be indifferent between several "hot" releases. Using this additional flexibility allows the firm to lower stock levels. We study this benefit under the simplifying assumption that all the substitutes are simultaneously released in §6.

To the best of our knowledge, this paper is the first to provide an analysis of the Netflix model. A recent paper, Randhawa and Kumar (2008), does consider a similar setting. However, it builds an On-Off source-based model for customers subscribing to a rental service and studies the firm's decision to offer a pay-per-use or subscription contract. The video rental industry in general has been the subject of a lot of interesting research. For example, Mortimer (2008) utilizes data collected at a large number (6,137) of video rental stores in the United States between 1998 and 2000 to compare the stocking levels, rental prices, etc. A regression analysis is performed to examine the effect of a revenue-sharing scheme on the retailer's profit. The analysis shows that the revenue-sharing scheme has a small positive effect on the retailer's profit for popular titles, and a small negative effect for less-popular titles. Tang and Deo (2008) study the competition between retailers on rental

price and rental duration instead of product availability. Kiesmüller and van der Lann (2001) studies dependency between the demand and return processes for the case of managing the inventory of a single type of a reusable product; purchasing lead time is also modeled. This paper demonstrates the extent of error incurred by neglecting the dependency between return process and the demand.

1.1. Notation

All random elements in this paper are defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Further, we assume all stochastic processes lie in the space of functions that are right continuous and possess left limits. For a collection of probability measures P^n and P defined on $(S, \mathcal{S})$, where S is a general metric space and $\mathcal{S}$ is its Borel σ-field, we say that as $n \to \infty$, $P^n \Rightarrow P$, i.e., P^n weakly converges to P, if and only if $\int_S f\,dP^n \to \int_S f\,dP$ for all bounded, continuous real-valued functions f on S. Further, if X^n and X are random elements of this space such that P^n and P are the probability measures associated with X^n and X, respectively, then $X^n \Rightarrow X$ if and only if $P^n \Rightarrow P$.

For functions $f, g\colon \mathbb{R} \to \mathbb{R}$, we shall say that $f(n) = O(g(n))$ if there exist constants $C_1, C_2 > 0$ such that $C_1 \leqslant f(n)/g(n) \leqslant C_2$ for all n. Further, we shall say that $f(n) = o(g(n))$ if $\lim_{n\to\infty}(f(n)/g(n)) = 0$.

2. Model

2.1. Setup and Assumptions

We begin by considering a system with n customers, indexed by $i = 1, \ldots, n$. We assume that customer i is of type $\theta \in \Theta$ with probability p_θ, where Θ is a finite set. We use θ_i to denote the type of customer i. We associate each customer type $\theta \in \Theta$ with a rental duration distribution F_θ that has support on $\mathbb{R}_+$, finite mean, and does not charge the origin, i.e., $F_\theta(0) = 0$. Let m_θ denote the mean of the distribution F_θ. As described in the introduction, in our model customers always have one product with them at any given time, asking for the next product upon the return of the previous rental. Even when they do not get their preferred product, we assume that they are given some product to rent, a perfectly reasonable assumption in systems like Netflix. Customer i holds onto her jth rental for a random time v_{ij}. We assume that $\{v_{ij}\}_{j=1}^{\infty}$ is a sequence of independent and identically distributed random variables, distributed according to the cumulative distribution function F_{θ_i}. In particular, the rental durations do not depend on the product being rented. We assume that the initial rental duration v_{i0} is a random variable that is independent of v_{ij} for $j = 1, 2, \ldots$. We will discuss the distribution of v_{i0} shortly. The rental times are assumed to be independent across customers.

At any given time t, the residual time $R_i(t)$ of customer i represents the remaining time left on her current rental. Because she obtains the next product only on the return of her current rental, her request for the next product occurs at time $t + R_i(t)$. Let $N_i(t)$ denote the counting process that counts the number of products rented by customer i by time t. That is, $N_i(t) = \sup\{j\colon \sum_{k=0}^{j} v_{ik} \leqslant t\}$, where we use the convention that $\sum_{k=0}^{j} v_{ik} = 0$ for $j < 0$. Let $T_i(\ell)$ denote the time instant at which the ℓth rental of customer i began, i.e., $T_i(\ell) = \sum_{j=0}^{\ell-1} v_{ij}$. We can write out customer i's residual rental time as

$$R_i(t) = v_{iN_i(t)} + T_i(N_i(t)) - t.$$

Now, suppose that at some arbitrary time T the new product is introduced. We assume that each customer wants the new product upon returning the product she is currently renting. As described in the introduction, we are interested in the dynamics of usage of this new product: The preceding assumption makes the analysis convenient by simply allowing us to ignore those people who do not wish to rent the new product. If the new product is available, it is given to the customer; otherwise, the customer rents an old product, which is defined to be any product other than the new one, and requests the new product again upon returning the old product. As mentioned earlier, customer i rents the new product for a duration that is identical in distribution to that of the old product. We further assume that the rental duration of the product the customer is renting when the new product is introduced is not affected by the introduction of the new product. Finally, we assume that customers do not rent the new product more than once.

The following result, which follows from §§2–16 in Wolff (1989), states the connection between the (stationary) residual rental time of customers and the excess distribution of F_θ. This allows us to make a convenient assumption that frees our analysis from dependence on the introduction time T.

PROPOSITION 1. *For each customer i, $i = 1, \ldots, n$ and $x \geqslant 0$, we have*:

1. (*Stationarity*). *If* $\mathbb{P}(v_{i0} > x \mid \theta_i = \theta) = (1/m_\theta)\int_x^\infty [1 - F_\theta(s)]\,ds$, *then* $\mathbb{P}(R_i(t) > x \mid \theta_i = \theta) = (1/m_\theta)\int_x^\infty [1 - F_\theta(s)]\,ds$ *for all* $t > 0$.

2. (*Steady state, time average*). *If v_{i0} has a finite mean, we obtain*

$$\lim_{t\to\infty} \frac{\int_0^t \mathbb{P}(R_i(s) > x \mid \theta_i = \theta)\,ds}{t} = \frac{\int_x^\infty [1 - F_\theta(s)]\,ds}{m_\theta} \quad \textit{and} \quad (1)$$

$$\lim_{t\to\infty} \frac{\int_0^t \mathbb{P}(R_i(s) > x)\,ds}{t} = \sum_{\theta\in\Theta} p_\theta \frac{\int_x^\infty [1 - F_\theta(s)]\,ds}{m_\theta}. \quad (2)$$

3. (*Steady state*). *If v_{i0} has a finite mean and F_θ is nonlattice for $\theta \in \Theta$, then we obtain*

$$\lim_{t\to\infty} \mathbb{P}(R_i(t) > x \mid \theta_i = \theta) = \frac{1}{m_\theta}\int_x^\infty [1 - F_\theta(s)]\,ds \quad \textit{and} \quad (3)$$

$$\lim_{t\to\infty} \mathbb{P}(R_i(t) > x) = \sum_{\theta\in\Theta} p_\theta \frac{\int_x^\infty [1 - F_\theta(s)]\,ds}{m_\theta}. \quad (4)$$

Note that (2) and (4) follow from (1) and (3) by taking an expectation on the customer type.

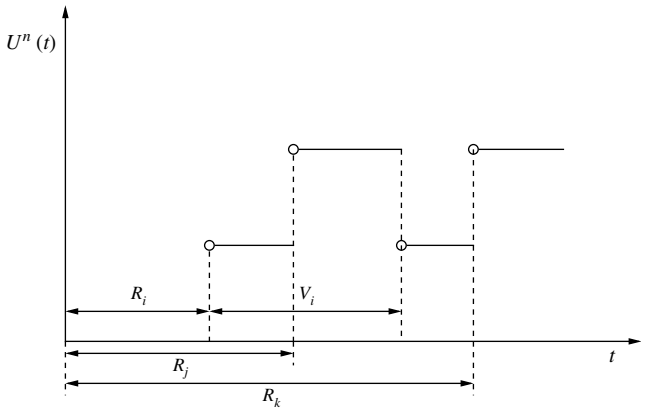
For any finite introduction time T, the behavior of the system will depend on T. In contexts such as Netflix, it is reasonable to assume that T is very large, i.e., the system has been operating for a long time when the new product is introduced, and in this case the dependence on the actual value of T can be ignored because the system is close to steady state. The proposition above justifies making the following assumption that conveniently allows us to ignore the actual value of T. We assume that v_{i0} is distributed according to the excess distribution of F_{θ_i} given by $F_{\theta_i e} \equiv (\int_0^x [1-F_{\theta_i}(s)]\,ds)/m_{\theta_i}$. Then, using Proposition 1, we conclude that the residual rental duration for the customers is an i.i.d. sequence $R_i(T)$ that is independent of the time of introduction T, and thus can be denoted $R_i(T) \stackrel{d}{=} R_i$, where R_i has the distribution $F_{\theta_i e}$. In addition, the duration of customer i's rental of the new product has the same distribution as v_{i1}. We shall henceforth use V_i to denote customer i's rental duration of the new product. Because the introduction time T becomes redundant for our analysis, we simply drop it from our notation. Finally, we note in passing that stationarity is different from steady state for the case when F_θ is a lattice distribution for any type θ—that is, there exists a d such that $\sum_{k=1}^{\infty} \mathbb{P}(v_{i1} = kd \mid \theta_i = \theta) = 1$. However, a Cesaro limit as in part 2 of the result holds, and we could use this as a proxy for the steady state. In any case, the process defined via our choice of v_{i0} is guaranteed to be stationary.

2.2. Dynamics of Usage

Suppose that the system manager stocks n units of the new product. Of course, this guarantees that no request for the new product will be denied. However, because the products are reusable, we expect the actual usage to be below n. That is, stocking n units is unnecessarily conservative. However, this will be the first step in our analysis. By quantifying the unconstrained dynamics of usage, we will be able to estimate the peak usage, and this provides a useful connection between rental distributions and stocking decisions.

Recall from the last subsection that the set of independent random pairs $\{R_i, V_i\}_{i=1}^n$ represents the request times and rental durations for the new product. These random variables will be used to write out the usage process $U^n(t)$ that tracks the number of units of the new product rented out at time t, where t measures time beyond the arbitrary time of introduction. (Note that the superscript n indicates the fact that there are n customers and n copies stocked.) It is worth pointing out that this usage process eventually gets absorbed at zero because each customer rents the product at most once, i.e., we have $U^n(0) = 0$ and $\lim_{t\to\infty} U^n(t) = 0$, and it is the dynamics of $U^n(t)$ over $t \in (0, \infty)$ in which we are interested. Figure 1 illustrates a sample path of the usage process. At any given time t, the customers who have a unit of the new product must

Figure 1. The usage process.



have (i) returned their previous rental by t, i.e., have $R_i \leqslant t$; (ii) must not have returned the new rental, i.e., have $R_i + V_i > t$. Those that have $R_i + V_i \leqslant t$ have returned the new rental and therefore no longer contribute to the usage; those with $R_i > t$ will only contribute to usage at some future time beyond t. Formally,

$$U^n(t) = \sum_{i=1}^{n} \mathbb{I}\{R_i \leqslant t \text{ and } R_i + V_i > t\}, \tag{5}$$

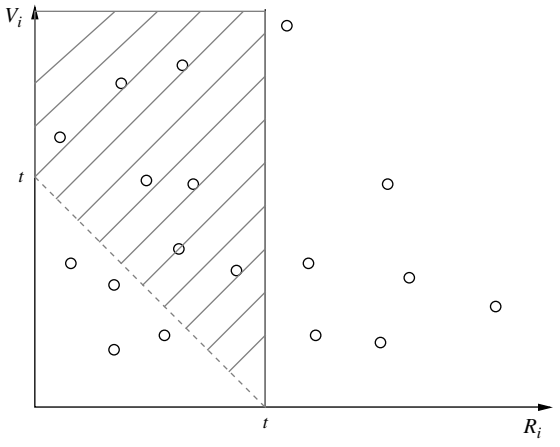
where $\mathbb{I}\{A\}$ is the indicator function of the event A.

We shall use a different interpretation of the usage process for our analysis. We can think of $\{R_i, V_i\}_{i=1}^n$ as being n i.i.d. draws from the distribution

$$\mathbb{G}(r, v) = \sum_{\theta\in\Theta} p_\theta F_{\theta e}(r) F_\theta(v).$$

Consider the empirical distribution corresponding to $\mathbb{G}$ constructed from the realization $\{R_i, V_i\}_{i=1}^n$. Figure 2 provides a graphical representation of this empirical distribution, with each atom denoting a customer with her residual time and rental duration. Then, according to (5), the number of atoms in the shaded region denotes $U^n(t)$. Note that based on this

Figure 2. An alternate representation.



representation, $U^n(t)$ is distributed as a binomial random variable for each $t \geqslant 0$ with n trials and probability of success $\mathbb{P}(R_i \leqslant t, R_i + V_i > t)$, and thus one can analyze the mean behavior $\mathbb{E}U^n(t) = n\mathbb{P}(R_i \leqslant t, R_i + V_i > t)$ with ease. However, the knowledge of this distribution at any fixed t gives no indication of the temporal behavior, and in particular, the peak of the usage process. This motivates us to consider asymptotic methods and analyze the system with a large number of customers. This is the natural asymptotic regime in applications like Netflix. In this setting, the simplicity of the representation in (5) allows us to invoke results from empirical process theory to obtain the limiting processes. Further, by characterizing the distribution of the usage process, we are able to obtain structural insights that will serve useful in computing stocking levels (see §5).

3. Asymptotic Analysis

As discussed above, we study asymptotic behavior of the system when the number of customers n grows without bound. (This limiting regime is akin to the Halfin-Whitt multiserver asymptotic regime.) As before, let $\mathbb{G}$ denote the probability measure associated with (R_i, V_i). Then, $U^n(\cdot)/n$ given by (5) is an empirical process associated with $\mathbb{E}\mathbb{I}\{R_i \leqslant \cdot, R_i + V_i > \cdot\}$ defined on $(D[0,\infty), d)$, where d is the metric as in §3.5 of Ethier and Kurtz (1986), which induces the Skorohod topology. Defining $u(t) \equiv \mathbb{E}\mathbb{I}\{R_i \leqslant t, R_i + V_i > t\} = \mathbb{P}(R_i \leqslant t, R_i + V_i > t)$ for $t \geqslant 0$, we can now apply the appropriate versions of the Glivenko-Cantelli and Donsker theorems (cf. Chapter 19 in van der Vaart 2000) to obtain the following asymptotic characterization of the usage process.

THEOREM 1. *As $n \to \infty$, we have*

1. (*Glivenko-Cantelli*)

$$\sup_{t \geqslant 0} \left| \frac{U^n(t)}{n} - u(t) \right| \to 0, \ \textit{a.s.} \tag{6}$$

2. (*Donsker*)

$$\sqrt{n}\left(\frac{U^n(\cdot)}{n} - u(\cdot) \right) \ \Rightarrow \ W(\cdot), \tag{7}$$

where $W(\cdot)$ is a continuous, zero-mean Gaussian process with covariance function $\gamma(s,t) = \mathbb{P}(R_1 \leqslant s \wedge t, R_1 + V_1 > s \vee t) - u(s)u(t)$.

These results hold for both discrete and continuous rental duration distributions. We postpone all proofs to the appendix.

The process $u(\cdot)$ characterizes the fraction of customers renting the new product as a function of time in a large system, and in this sense describes the usage of the product asymptotically. Note that if we were to fix any $t > 0$, the asymptotic properties of $U^n(t)/n$ can be calculated by a direct application of the strong law of large numbers and the central limit theorem. However, this pointwise convergence does not characterize the process sufficiently to identify the peak of the usage process. To do this, establishing process-level convergence as in Theorem 1 is essential. Using this convergence, we show in Corollary 1 below that the peak of the usage process is asymptotically equal to the peak of the process $u(\cdot)$, b^*, defined as follows:

$$\begin{aligned} b^* &\equiv \sup_{t \geqslant 0} \mathbb{P}(R_i \leqslant t, R_i + V_i > t) \\ &= \sup_{t \geqslant 0} \sum_{\theta \in \Theta} p_\theta \mathbb{P}(R_i \leqslant t, R_i + V_i > t \mid \theta_i = \theta) \\ &\overset{(a)}{=} \sup_{t \geqslant 0} \sum_{\theta \in \Theta} p_\theta \int_0^t \mathbb{P}(V_i > t - s)\, dF_{\theta e}(s) \\ &= \sup_{t \geqslant 0} \sum_{\theta \in \Theta} p_\theta \frac{\int_0^t (1 - F_\theta(s))(1 - F_\theta(t-s))\, ds}{m_\theta}, \end{aligned} \tag{8}$$

where (a) follows by the distribution function of R_i when $\theta_i = \theta$ is given by $F_{\theta e}$. Now, taking the supremum over $\{t \geqslant 0\}$ in (6)–(7) and arguing the interchange of the limit and the supremum function, we obtain the following result, which characterizes the peak of the usage process. (This is proved in the appendix.)

COROLLARY 1. *As $n \to \infty$, we have*

1. $\sup_{t \geqslant 0}(U^n(t)/n) \to b^*$, *a.s.*
2. $\sqrt{n}(\sup_{t \geqslant 0}(U^n(t)/n) - b^*) \Rightarrow \sup_{s \in S} W(s)$, *where* $S = \{t\colon u(t) = b^*\}$.

This result implies that for large n, the peak of the usage process can be loosely written as

$$\sup_{t \geqslant 0} U^n(t) = b^* n + \sup_{s \in S} W(s)\sqrt{n} + O(\sqrt{n}).$$

Thus, for large n, the peak of the usage process will be governed by the term b^*, which in turn is the peak of the mean usage process $u(\cdot)$. Before proceeding, we illustrate the behavior of $u(\cdot)$ via numerical examples.

Figures 3 and 4 plot the mean usage process $u(\cdot)$ for a homogeneous population with rental durations distributed according to a unit mean exponential distribution and a

Figure 3. $u(\cdot)$: Exponential rental distribution.

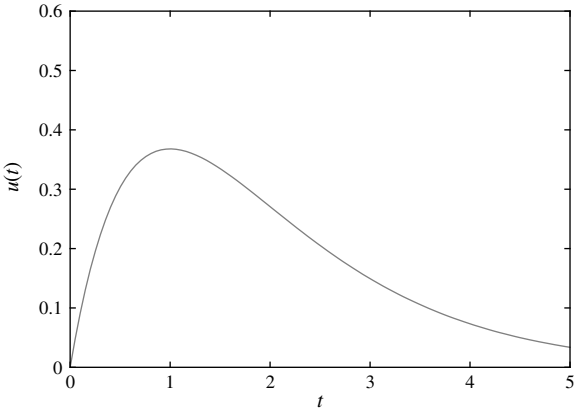
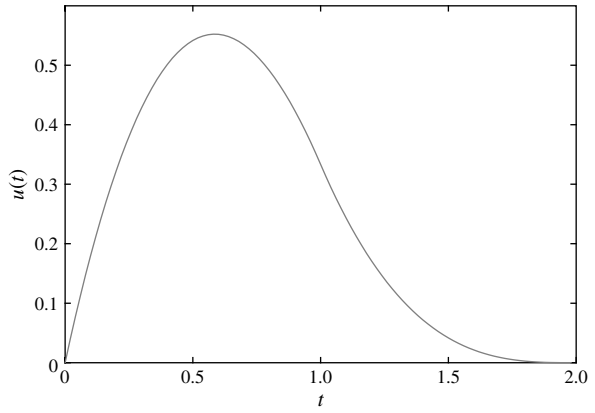


Figure 4. $u(\cdot)$: Uniform $[0, 1]$ rental distribution.



uniform distribution over $[0, 1]$, respectively. It is possible to have rental distributions that are not as well-behaved as the ones in these figures. Namely, $u(\cdot)$ is not always unimodal. One can have rental distributions for which $u(\cdot)$ has multiple local maximas or achieves its maximum over an interval. The behavior of $u(\cdot)$ depends on the entire distribution of the rental duration, and not on the first or second moment alone. This motivates the study of the relationship between the rental duration distributions F_θ and the peak of the mean usage process, b^*. As the reader might guess, b^* is a crucial component of any stocking decision (we discuss this connection in §5).

4. Peak Mean Usage and Rental Distributions

Our aim in this section is to use the asymptotic analysis of the previous section to develop some structural insights about b^*. First, we shall characterize the *ideal* customer population. To be precise, we will try to find the duration distribution with a given mean that has the lowest b^*. Unlike traditional inventory management problems, it is not true that the deterministic distribution is ideal. In fact, this distribution turns out to be the worst (see Proposition 2). Section 4.1 shows that there is an ideal customer population in a limiting sense—that is, b^* can be made arbitrarily small for any given mean. In §4.2, we demonstrate the invariance property of the exponential distribution with respect to b^*, namely, $b^* = e^{-1}$ regardless of the mean rental duration. Next, in §4.3, we investigate the impact of imposing rental deadlines on the peak usage. We show that the longer (i.e., more relaxed) the deadlines, the lower the b^*. Finally, in §4.4, we show that approximating a heterogeneous customer population by an equivalent homogeneous one can lead to an arbitrarily large error.

4.1. The Ideal Customer Population

In this subsection, we assume all customers to be homogenous, i.e., $F_\theta = F$ for all $\theta \in \Theta$. Our aim is to perform a best- and worst-case analysis of b^* over rental distributions that have a given mean. It is useful to define some notation first. Denote the set of all distributions with mean $m > 0$ by $\mathcal{F}_m$, and for any rental time distribution $F \in \mathcal{F}_m$, denote the peak of mean usage by b_F^*. Equation (8) implies that $b_F^* \leqslant 1$ for all rental distributions and for a deterministic distribution with any mean m, the corresponding $b^* = 1$. Thus, we obtain the following result that we state without proof.

Proposition 2 (Worst Rental Distribution). *A deterministic rental distribution $F(x) = 1_{\{x \geqslant m\}}$ has the largest b_F^*, namely, unity.*

We now characterize the best rental distribution.

Proposition 3 (Ideal Rental Distribution). 1. *For any $\epsilon \in (0, m)$, the rental distribution $F_\epsilon \in \mathcal{F}_m$ given by*

$$F_\epsilon(x) = \begin{cases} 0, & x < \epsilon, \\ 1-\epsilon, & \epsilon \leqslant x < m/\epsilon + \epsilon - 1, \\ 1, & \text{otherwise}, \end{cases} \tag{9}$$

has $b_{F_\epsilon}^* = O(\epsilon)$. Further, the second moment

$$\int_0^\infty x^2 \, dF_\epsilon(x) = O(1/\epsilon).$$

2. $\inf_{F \in \mathcal{F}_m} b_F^* = 0$.

Proposition 3 shows that it is possible to make b^* to be $O(\epsilon)$ for any $\epsilon > 0$. That is, it is possible to get distributions that are arbitrarily good, even though there is no single "best" rental distribution or ideal population. The way in which b^* is made small results in the variance of the distribution becoming arbitrarily large. Although it is usually true in stochastic systems that increasing variance degrades performance, in this special case, the opposite is true in our system. The prescriptive implications of this result are not immediately apparent for a firm like Netflix: Inducing such rental behavior among rational customers appears to be an interesting topic for future work.

An intuitive explanation of this result is as follows. The Inspection Paradox (cf. §§2–4, Wolff 1989) tells us that at the time of introduction of the new product, customers (who have distribution F_ϵ) are disproportionately likely to be in their longer rental duration, which is of size $O(1/\epsilon)$, and therefore the residual time of the customers will be uniformly spread over an interval of size $O(1/\epsilon)$. That is, requests for the new product come evenly spread over a long interval. However, on receiving the new product customers are more likely to return the product within $O(\epsilon)$. This combination of spread-out requests followed by quick returns results in a unit of the new product being reused $O(1/\epsilon)$ times, and therefore $u(\cdot)$ does not get above a level that is $O(\epsilon)$.

4.2. The Exponential Distribution: $b^* = e^{-1}$

It is easy to see that the peak usage level is independent of any time scaling, i.e., $\sup_{t \geqslant 0} U^n(t) = \sup_{t \geqslant 0} U^n(\gamma t)$ for

any $\gamma > 0$. We illustrate the asymptotic version of this independence for the case of exponential distributions. Because a rescaled exponential distribution is also exponential, this leads us to the following mean invariance result.

PROPOSITION 4. *For a homogeneous population with exponential rental distributions, i.e., $F(x) = 1 - e^{-x/m}$ for $x \geqslant 0$, we have $b^* = e^{-1}$ for any $m > 0$.*

This result, when combined with Proposition 3, allows us to also conclude that b^* is not monotone in the mean m. In particular, given a mean m, one can find a two-point distribution that has b^* smaller than e^{-1}. Of course, the deterministic distribution with the same m has $b^* = 1$. Furthermore, $u(\cdot)$ constructed for the uniform distribution in Figure 4 allows us to show that b^* is not monotone in the second moment either because we get $b^* = e^{-1}$ for the exponential distribution regardless of its variance.

4.3. Effect of Deadlines: Longer Deadlines Imply Lower b^*

It is common for most rental firms to impose deadlines on customer rentals. A firm wanting to increase reuse may choose to impose return deadlines to ensure that a unit of the desired new product is not kept by a user for a long time. On the flip side, imposing deadlines makes sure that the customers return old products quicker as well, and therefore end up asking for the new product earlier than they might have without deadlines. It is not a priori clear which way this trade-off is resolved with regard to b^*. Netflix chooses not to impose any deadlines on its customers. This could be for various reasons, such as strategic differentiation from their main competitor Blockbuster, or to ensure that the mean rental time is not reduced, resulting in higher postage and other transaction costs per unit time. The question we address is whether these benefits of not imposing deadlines could be offset by higher stocking costs resulting from a higher b^*.

To do so, we first need a model of how customers react to deadlines. We assume that the firm imposes a deadline d on all rentals and all customers always return the product before (or at) the deadline. As before, we assume that each customer has an underlying rental distribution F_θ from which the amount of time the customer *wishes* to rent is drawn. This time can be thought of as the customer's rental duration when there are no deadlines. When a deadline of d is imposed, the customer will only rent for a duration that is the minimum of her wished rental time and the deadline duration d. Thus, in our model, the cumulative distribution of the rental time with a deadline of d is given by

$$F_\theta^d(x) = \begin{cases} F_\theta(x), & x < d, \\ 1, & x \geqslant d. \end{cases}$$

Denoting the peak of mean usage by b_d^*, we obtain

$$b_d^* = \sup_{0 \leqslant t \leqslant 2d} \sum_{\theta \in \Theta} p_\theta \frac{\int_0^t (1 - F_\theta^d(s))(1 - F_\theta^d(t-s))\,ds)}{\int_0^d (1 - F_\theta^d(s))\,ds}. \quad (10)$$

Then, under our proposed model of deadlines, we have the following result.

PROPOSITION 5. *The peak of the mean usage decreases as the imposed deadline becomes more relaxed, i.e., b_d^* is decreasing in d.*

The mean of the distribution F_θ^d is increasing in d. However, given the conclusion from the previous subsection that b^* is not monotone in the mean, this is not sufficient to conclude that b^* is decreasing with d. Proposition 5 could not have been obtained without considering the actual dynamics of $u(\cdot)$. Indeed, this is how the proof of Proposition 5 proceeds in the appendix.

4.4. Approximating a Heterogeneous Pool by Homogeneous Customers

Our analysis in the previous two sections allowed for a heterogeneous customer population. The treatment of heterogeneity is not mathematical generality for its own sake. In particular, heterogeneity in the duration distributions cannot be modeled away by approximating the heterogeneous population with an equivalent homogeneous population whose rental duration distribution is an appropriate mixture of the constituent duration distributions. Via an example, we demonstrate that it is possible to make an arbitrarily large error in calculating b^* doing this. This has obvious implications for empirical estimation. It is important to segregate customers into types when estimating duration distributions.

Consider a customer population consisting of two types θ_1, with a deterministic rental duration of ϵ, and θ_2, with a deterministic rental duration of $m/\epsilon + \epsilon - 1$. Suppose that $p_{\theta_1} = 1 - \epsilon$ and $p_{\theta_2} = \epsilon$. Using (8), we see that b^* for this population is at least $1 - \epsilon$. The equivalent homogeneous model for this population is the two-point rental distribution given in (9) with a b^* that is $O(\epsilon)$. Clearly, the magnitude of the error in this approximation can be made arbitrarily large. The explanation for this effect is simple. What goes into calculating b^* is $\mathbb{G}$, which involves a mixture of residual time distributions. Of course, this need not be distributed the same as the residual time corresponding to the mixture of duration distributions.

5. Stocking

Thus far, we have investigated the peak of the usage process for the customers in the setting where sufficiently many copies of the new product are stocked upon introduction. A natural subsequent task is to determine the optimal number of copies of the new products to stock. To do this, we could build an economic framework assigning costs to stocking the products and to denied customer requests. Alternatively, we can define a performance measure and find the smallest stock level that achieves an acceptable level of this performance measure. We choose the latter approach. Two performance measures that one can envisage

are (i) probability that a stockout occurs, henceforth called stockout probability; or (ii) the denial count, which measures the number of requests for new products that are denied.

We first deal with the stockout probability. Let b^n denote the number of units of the new product stocked in the system with n customers, and let $\alpha^n(b^n)$ denote the resulting stockout probability. The observation that the usage process for any stock level is identical to the unrestricted usage process up until the first time a stockout occurs allows us to compute the stockout probability as

$$\alpha^n(b^n) = \mathbb{P}\Big(\sup_{t\geqslant 0} U^n(t) > b^n\Big). \tag{11}$$

This relation allows us to utilize the asymptotic analysis of U^n carried out in the previous sections.

We are interested in finding the smallest b^n that ensures that the consequent stockout probability $\alpha^n(b^n)$ is no bigger than some target $\bar{\alpha}$. As before, we will answer this in the natural asymptotic regime where n grows without bound. The following result, which is a consequence of Corollary 1, characterizes the asymptotically minimal stock level b^n given an $\bar{\alpha}$.

Proposition 6. *For a sequence of stock levels* $\{b^n\colon n = 1, 2, \ldots\}$, *if* $(b^n - nb^*)/\sqrt{n} \to \hat{b} < \infty$ *as* $n \to \infty$, *then* $\alpha^n(b^n) \to \mathbb{P}(\sup_{s\in S} W(s) > \hat{b})$ *as* $n \to \infty$. *Further, if* $(b^n - nb^*)/\sqrt{n} \to \infty$ *as* $n \to \infty$, *then* $\alpha^n(b^n) \to 0$ *as* $n \to \infty$.

As may be expected, Proposition 6 says that b^n should be nb^* on the $O(n)$ scale. Of course, this need not be true in the trivial case when $\bar{\alpha} = 1$. For a given performance level $\alpha \in (0, 1)$, if one can compute $\hat{b}$ such that $\mathbb{P}(\sup_{s\in S} W(s) > \hat{b}) = \alpha$, then this result implies that stocking at a level $b^n = nb^* + \sqrt{n}\hat{b}$ meets this performance criterion. Further, we can guarantee near-perfect service with a stock level that is only slightly bigger than nb^*. In fact, one can refine the excess stock level above nb^* to any level $f(n)$ such that $f(n)/\sqrt{n} \to \infty$. Therefore, b^* is what really matters in this problem.

When $\bar{\alpha} \in (0, 1)$, computing the $O(\sqrt{n})$ refinement of the corresponding stock level requires calculating the distribution of $\sup_{t\in S} W(t)$. When the set S is a singleton (as it is for the exponential and uniform distributions depicted in Figures 3 and 4), $\sup_{t\in S} W(t) = W(t^*)$, which is a Normal random variable. In this case, computing $\hat{b}$ is straightforward. When S is not a singleton but is a finite set, we need to compute the probability that the maximum component of a multivariate Normal random variable exceeds $\hat{b}$. This is still tractable. The case when S is a general uncountable set is a classical open problem. There are asymptotic results when $\bar{\alpha}$ is small; we refer the reader to Adler and Taylor (2007).

We now turn our attention to the second performance measure, the denial count. Unfortunately, we are unable to characterize the denials as cleanly as the stockout probability. For computing the denials, one needs to keep track of the retrial process of customers denied the new product in a previous attempt. This renders the analysis considerably more complex. However, we can use the asymptotic analysis in §3 to characterize stock levels that lead to negligible denials. Denoting the total number of requests for the new product denied as a function of the capacity level by $d^n(b^n)$, we obtain the following result.

Proposition 7. *The denial count* $d^n(b^n) \to 0$ *a.s. as* $n \to \infty$ *for the stock level* $b^n = (b^* + \epsilon)n$ *for any* $\epsilon > 0$. *Further, for almost all* $\omega \in \Omega$, *there exists* $N(\omega) < \infty$ *such that* $d^n(b^n, \omega) = 0$ *for* $n > N(\omega)$.

Obtaining the general version of Proposition 7 when the desired denial count is nonnegligible, i.e., $d^n(b^n) = O(n)$, appears to be a challenging problem. In particular, tools beyond those developed in this paper may prove necessary. We leave this as a topic for future work.

So far, we have focused on the introduction of a single new product. However, it may be the case that several new products are introduced within a short time span. This is definitely true in the DVD world. Clearly, the introduction of multiple products has an implication on the usage process of the customers, and thus on the firm's stocking decisions. In the following section, we demonstrate how we can apply the tools we have developed to this setting.

6. Multiple Product Introduction

We consider the setting where k distinct new products are introduced simultaneously at some arbitrary time. We assume that customers request each of these k products before requesting any other product. However, at the time of request, the customer is indifferent between any of the newly introduced products that the customer has not yet rented. The system manager is free to exploit this indifference. We retain the assumptions on the rental duration distributions made in §2, although for simplicity we will restrict our attention to the case of homogeneous customers.

To completely characterize the dynamics of this system, we need to keep track of the usage of each individual product, as well as the set of customers who have already rented specific subsets of the introduced products. However, characterizing the total usage of all these k products is simpler. Customers who are renting one of the k new products at time t have a residual time at the time of introduction smaller than t and the sum of the k subsequent rental durations larger than t. (We do not know which of the k products each of these customers have at t, just that they have one.) Denote the usage process for each of the new products by $U_j^n(\cdot)$ for $j = 1, 2, \ldots, k$. Defining $\bar{u}(t) \equiv \mathbb{P}(R_1 \leqslant t, R_1 + \sum_{j=1}^k V_{1j} > t)$, we can argue as in §3 to asymptotically characterize the total usage process in the following fashion.

Proposition 8. $\sup_{t\geqslant 0} |(1/n)\sum_{j=1}^k U_j^n(\cdot) - \bar{u}(\cdot)| \to 0$ *a.s., as* $n \to \infty$.

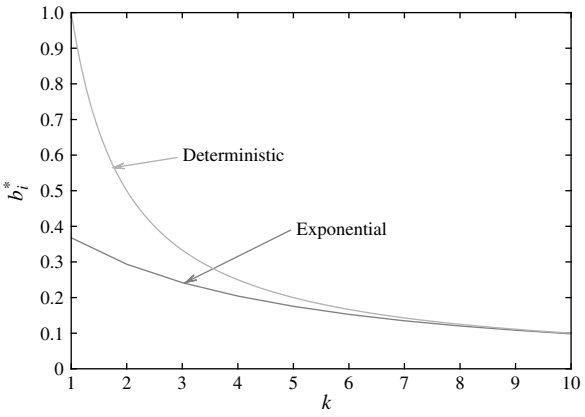
This result is an immediate extension of Theorem 1.1, and we omit its proof. The peak of the total usage $\overline{U}^n \equiv \sum_{j=1}^k \sup_{t\geqslant 0} U_j^n(t)$ is a measure of the overall stock level required in this setting along the lines of the discussion in §5. At any point in time, the manager would like to use the customers' indifference between the various new products to rent out products in a way that would lead to lower $\overline{U}^n$ values. Thus, in addition to total usage, we also need to compute a dynamic product allocation policy. However, when implementing an optimal policy for any symmetric cost criterion, we would want the usage of the products to be equal to each other as much as possible. Therefore, the corresponding asymptotic peak usage level is given by

$$b_i^* \equiv \frac{1}{k}\lim_{n\to\infty}\frac{\overline{U}^n}{n} = \frac{1}{k}\sup_{t\geqslant 0}\bar{u}(t).$$

In passing, we construct a policy that asymptotically achieves this individual peak usage level for each product. To ensure a symmetric usage, one can create $k!$ classes of customers, where each class corresponds to a unique permutation of the order in which the products will be rented. We assign customers to each of these classes in a uniform manner. That is, each class will have $n/k!$ customers (where we drop the integrality requirement for convenience). The allocation of the products is as follows: Upon requesting a product for the first time after introduction, each customer is given the first product in the permutation that corresponds to their class. Upon return of this rental, the customer is given the second product in their permutation, and future allocations proceed in the same fashion until customers have rented all the new products. This allocation ensures that the usage of all products will be identical. For example, consider this policy for the case of $k = 2$ products. Here, we divide the customer population into two classes. The first class corresponds to customers who rent product 1 before product 2. Similarly, the second class corresponds to customers who rent product 2 first. By assigning half the customer population to each class, we ensure an identical usage of both the products, and hence obtain the desired symmetry. Note that this policy is asymptotically equivalent to a randomized policy, where upon request by a customer, she is allocated a product selected randomly (uniformly) from the set of products she has not yet rented.

The scaled total usage process is always bounded above by one, i.e., $(1/n)\sum_{j=1}^k U_j^n(\cdot) \leqslant 1$, and thus $\bar{u}(t) \leqslant 1$ for all $t \geqslant 0$. Further, $b_i^* \leqslant 1/k$ for all $i = 1, \ldots, k$ given our allocation policy. For a homogeneous population with a deterministic rental distribution, $\sup_{t\geqslant 0}\bar{u}(t) = 1$. Thus, $b_i^* = 1/k$. For a homogeneous population with an exponential rental distribution with mean m, $\bar{u}(t) = (\sum_{j=1}^k t^j/(m^j j!))e^{-t/m}$ and $b_i^* = (1/k)\sup_{t\geqslant 0}\bar{u}(t)$. Figure 5 plots the peak mean usage for one new product for a homogeneous population for the case of exponential and deterministic rental distributions as a function of the number of distinct new products introduced. Observe that for a small number of products, the exponential rental distribution has a lower peak usage, but this difference decreases as more and more products are released simultaneously. This suggests that when multiple products are introduced simultaneously, stocking decisions made using a deterministic rental distribution assumption can perform fairly well. That is, the so-called *multiplexing benefit* from variability in the rental distribution becomes less significant as the degree of substitution increases.

Figure 5. Exponential vs. deterministic rental distributions: Multiple products.



Finally, suppose that the new products are not introduced simultaneously. In this case, it is no longer true that $\overline{U}^n$ is bounded above by n. In fact, if the introduction times for the products are sufficiently staggered (on a scale measure by m_θ), the problem can be approximately solved using k replications of the analysis in §2. If the introduction times are close together, then the analysis in this section would approximately apply. Analysis of the general introduction problem is fairly involved and we leave it as a topic for future research.

7. Conclusion

In this paper, we develop a model for rental systems where each rental completion triggers a demand for another product. In the setting where each customer rents a particular product only once, which is the case of Netflix, our canonical rental firm, we study the introduction of a new product from a stocking perspective. Using classical empirical process theory, namely, the Glivenko-Cantelli and Donsker theorems, we characterize the asymptotic behavior of the number of copies of the new product being rented at any time. We note that the peak of this usage process has important implications for stocking with respect to a quality criterion. In particular, a "high-quality" system is possible only when the stock level of the new product is close to this asymptotic peak. We characterize this level up to a $O(\sqrt{n})$ level for the stockout criterion. The criterion of number of denied requests (related to fill rate) is more involved, and although we characterize stock levels where the denied

requests are asymptotically negligible, the case of nonzero denials is also important. Unfortunately, this case is not quite tractable and requires the analysis of measure-valued limits, which goes beyond the scope of this paper and merits future work. Note that the case of nonzero denials corresponds to a scarcity in capacity, and in this setting prioritizing customer classes to optimize a social objective via some form of dynamic capacity allocation may be needed. Bassamboo and Randhawa (2009) study the structure of the optimal control policy in such settings.

Our analysis assumes that customers can only rent one product at a time. This assumption can be easily relaxed as follows: If each customer rents k products at a time, then the request time for a newly released product will be the minimum residual time across the k products. The rest of the analysis proceeds identically, using this minimum in place of the random variable R.

We also consider the case of multiple product introduction. We analyze this setting when all products are introduced simultaneously, and the customers are indifferent between them. The general case where these assumptions are relaxed is also important, and will be a useful extension. The case when customers are allowed to rent multiple products simultaneously, although straightforward for introduction of a single product, becomes more complicated when studying multiple product introduction, and is worth exploring.

Appendix. Proofs

We begin by establishing the following measurability result:

LEMMA 1. *For each* $n = 1, 2, \ldots,$

1. U^n *is a measurable map from* $(\Omega, \mathcal{F})$ *to* $(D[0,\infty), \mathcal{D})$, *where* $D[0,\infty)$ *is endowed with the Skorohod topology and* $\mathcal{D}$ *is the Borel* σ*-field on* $D[0,\infty)$.

2. $\sup_{t>0} U^n(t)$ *is a measurable map from* $(\Omega, \mathcal{F})$ *to* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, *where* $\mathcal{B}(\mathbb{R})$ *denotes the Borel* σ*-field on* $\mathbb{R}$.

PROOF. 1. The mappings ψ: $\mathbb{R}\times\mathbb{R} \mapsto D[0,\infty)$ given by $\psi(x,y) = \mathbb{I}\{x \leqslant \cdot, x+y > \cdot\}$ are continuous in the Skorohod topology, and hence $U^n = \sum_{i=1}^n \psi(R_i, V_i)$ is measurable.

2. Noting that U^n is an RCLL (right continuous with left limits) function, $\sup_{t\geqslant 0} U^n(t)$ can be equivalently computed by restricting the supremum to the set of rationals, thus $\sup_{t\geqslant 0} U^n(t)$ is measurable (as in the discussion following Theorem 20.6 in Billingsley 1995).

PROOF OF THEOREM 1. We define the functions $\phi_t : \mathbb{R}_+^2 \mapsto \{0,1\}$ for all $t \geqslant 0$ as follows:

$$\phi_t(x,y) = \begin{cases} 1 & \text{if } x \leqslant t \text{ and } x+y > t, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $U^n(t)$ given by (5) can be written as $\sum_{i=1}^n \phi_t(R_i, V_i)$. Let Φ be the class of functions ϕ_t for all $t \geqslant 0$, i.e., $\Phi = \{\phi_t\colon t \geqslant 0\}$.

We shall use the results from empirical process theory to show the convergence of $U^n(\cdot)$ on the fluid and diffusion scales. For this, it suffices to show that the class of functions Φ is $\mathbb{G}$-Glivenko-Cantelli and $\mathbb{G}$-Donsker (see pp. 269–270 in van der Vaart 2000). We show that Φ can be covered by a finite number of ϵ-brackets and this number grows polynomially as ϵ shrinks to zero. The result then follows from Theorems 19.4 and 19.5 in van der Vaart (2000).

Given two functions l and u, the bracket $[l,u]$ is the set of all functions f with $l \leqslant f \leqslant u$. For a probability distribution $\mathbb{H}$, an ϵ-bracket in $L_r(\mathbb{H})$ is a bracket $[l,u]$ with $\mathbb{E}_{\mathbb{H}}[u-l]^r = \int (u-l)^r \, d\mathbb{H} < \epsilon^r$. The brackets $[l_j, u_j]$, $j = 1, \ldots, J$ cover Φ if $\Phi \subseteq \bigcup_{j=1}^J [l_j, u_j]$.

We begin by constructing the ϵ-brackets in $L_1(\mathbb{G})$ needed to cover the functions Φ. Choose $0 = t_0 < t_1 < \cdots t_K = \infty$ such that $\sum_{\theta\in\Theta} p_\theta F_{\theta e}(t_{k+1}) - \sum_{\theta\in\Theta} p_\theta F_{\theta e}(t_k) < \epsilon/2$ for each k. Further, choose $0 = s_0 < s_1 < s_2 < \cdots < s_L = \infty$ such that $\mathbb{P}(R_1 + V_1 \leqslant s_{\ell+1}) - \mathbb{P}(R_1 + V_1 \leqslant s_\ell) < \epsilon/2$ for each ℓ. Let $0 = v_0 \leqslant v_1 \leqslant v_2 \leqslant \ldots v_M = \infty$ with $M \leqslant K + L - 1$ denote the sorted list of the set $\{t_0, t_1, \ldots, t_K\} \cup \{s_0, s_1, \ldots, s_L\}$. Define the functions

$$\ell_j(x,y) = \mathbb{I}\{x \leqslant v_j, x+y > v_{j+1}\} \quad \text{and}$$

$$u_j(x,y) = \mathbb{I}\{x \leqslant v_{j+1}, x+y > v_j\}$$

for $j \leqslant K+L+1$. Note that $\mathbb{E}_{\mathbb{G}}[u_j - \ell_j] \leqslant \epsilon$; thus, $[\ell_j, u_j]$ is an ϵ-bracket in $L_1(\mathbb{G})$. The collection of brackets $[\ell_j, u_j]$ for $j < K+L+1$ covers the set of functions Φ. Noting that $F_{\theta e}$ is continuous (even if the rental duration distribution is discrete), we can choose points $t_0, t_1, \ldots, t_K$ such that $K \leqslant C/\epsilon$, where C is any constant strictly greater than two. Similarly, noting that $R_1 + V_1$ also has a continuous distribution function, we can choose points $s_0, s_1, \ldots, s_L$ such that $L \leqslant C/\epsilon$. Thus, the number of ϵ-brackets $L_1(\mathbb{G})$ needed to cover Φ, denoted by $N_{[\,]}(\epsilon, \Phi, L_1(\mathbb{G}))$, is finite for every $\epsilon > 0$. Using Theorem 19.4 from van der Vaart (2000), we then have that Φ is $\mathbb{G}$-Glivenko-Cantelli. Thus, we have

$$\lim_{n\to\infty} \sup_{t\geqslant 0} \left| \frac{U^n(t)}{n} - u(t) \right| = 0, \quad \text{a.s.}$$

Next, note that $\mathbb{E}_{\mathbb{G}}[u_j - \ell_j]^2 \leqslant \epsilon$. So, $[\ell_j, u_j]$ is also an $\sqrt{\epsilon}$-bracket in $L_2(\mathbb{G})$. Thus, we have that the bracketing number $N_{[\,]}(\sqrt{\epsilon}, \Phi, L_2(\mathbb{G})) \leqslant 2C/\epsilon$. (Note that we do not need any second-moment condition on the distribution because the functions in the set Φ are bounded by one.) In addition, we have that the bracketing integral

$$\begin{aligned} J_{[\,]}(1, \Phi, L_2(\mathbb{G})) &= \int_0^1 \sqrt{\log N_{[\,]}(\epsilon, \Phi, L_2(\mathbb{G}))}\, d\epsilon \\ &\leqslant \int_0^1 \sqrt{\log\left(\frac{2C}{\epsilon^2}\right)}\, d\epsilon \\ &= \int_0^\infty \sqrt{\log 2C + 2y}\, e^{-y}\, dy < \infty, \end{aligned}$$

where the last equality follows by substituting $y = -\log \epsilon$. Applying Theorem 19.5 of van der Vaart (2000), we have that Φ is $\mathbb{G}$-Donsker. Thus, we have

$$\frac{U^n(\cdot) - n\mathbb{P}(R_1 \leqslant \cdot, R_1 + V_1 > \cdot)}{\sqrt{n}} \Rightarrow W(\cdot), \quad \text{as } n \to \infty,$$

where $W(\cdot)$ is a $\mathbb{G}$-Brownian bridge. Note that using Lemma 18.15 from van der Vaart (2000), we have that W is a zero-mean Gaussian process. Using the discussion on p. 269 of van der Vaart (2000), for $s \leqslant t$ we can compute the covariance function γ as follows:

$$\begin{aligned}\gamma(s,t) &= \mathbb{E}[\phi_s(R_1,V_1)\phi_t(R_1,V_1)] \\ &\quad - \mathbb{E}[\phi_s(R_1,V_1)]\mathbb{E}[\phi_t(R_1,V_1)] \\ &= \mathbb{P}(R_1 \leqslant s, R_1 + V_1 > t) - u(s)u(t).\end{aligned}$$

Applying Lemma 18.15 in van der Vaart (2000), we obtain that $W(\cdot)$ is continuous with respect to the semimetric ρ defined as $\rho(s,t)^2 = \mathbb{E}(W(s) - W(t))^2 = u(s) + u(t) - 2\mathbb{P}(R_1 \leqslant s, R_1 + V_1 > t)$. Noting the continuity of the distribution function of R_1, we obtain that for any sequence $\{t^r \in \mathbb{R}_+ \colon r = 1, 2, \ldots\}$, if $t^r \to t$ as $r \to \infty$, then $\rho(t^r, t) \to 0$. Thus, $W(t^r) \to W(t)$ a.s., as $r \to \infty$, and so W is continuous a.s. □

Proof of Corollary 1. Using Theorem 1.1 and the continuity of the supremum function with respect to the uniform metric on $D[0,\infty)$, we have

$$\sup_{t \geqslant 0} \frac{U^n(t)}{n} \to \sup_{t \geqslant 0} u(t) = b^* \qquad \text{a.s., as } n \to \infty. \tag{12}$$

We end the proof of part 1 of the result by showing that $b^* < \infty$ and u achieves its supremum. Note that R_1 is a finite-valued random variable with a continuous distribution. Thus, $u(t) = \mathbb{P}(R_1 \leqslant t, R_1 + V_1 > t)$ is continuous with a value of zero when $t = 0$ and approaches zero as $t \to \infty$. Thus, this function must achieve its supremum, i.e., there exists $t^* < \infty$ such that $u(t^*) = b^*$. Further, the set $S \equiv \{t \colon u(t) = b^*\}$ is compact.

We now prove part 2. Applying the Skorohod representation theorem (see, for example, Theorem 1.8 in Chapter 3 of Ethier and Kurtz 1986, or Theorem 5.1 in Chen and Yao 2001), we obtain the existence of random variables $\widetilde{U}^n$ and $\widetilde{W}$ defined on a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ with $\widetilde{U}^n \stackrel{d}{=} U^n$ and $\widetilde{W} \stackrel{d}{=} W$ and

$$\sup_{0 \leqslant t \leqslant T} \left| \sqrt{n}\left(\frac{\widetilde{U}^n(t)}{n} - u(t) \right) - \widetilde{W}(t) \right| \to 0,$$
$$\widetilde{\mathbb{P}}\text{-a.s., as } n \to \infty \text{ for all } T > 0. \tag{13}$$

Thus, to complete the proof, it suffices to prove that

$$\lim_{n \to \infty} \sqrt{n}\left(\sup_{t \geqslant 0} \frac{\widetilde{U}^n(t)}{n} - b^* \right) = \sup_{s \in S} \widetilde{W}(s), \quad \widetilde{\mathbb{P}}\text{-a.s.}$$

We use a bounding argument to prove this statement. Note that we have

$$\frac{(\sup_{t \geqslant 0} \widetilde{U}^n(t) - nb^*)}{\sqrt{n}} \geqslant \sup_{s \in S} \frac{(\widetilde{U}^n(s) - nb^*)}{\sqrt{n}}.$$

Because S is compact, for any two real-valued functions f, g, we have $|\sup_{s \in S} f(s) - \sup_{s \in S} g(s)| \leqslant \sup_{s \in S} |f(s) - g(s)|$. Combining this with the convergence in (13), we obtain

$$\left| \sup_{s \in S} \frac{(\widetilde{U}^n(s) - nu(s))}{\sqrt{n}} - \sup_{s \in S} \widetilde{W}(s) \right| \to 0$$
$$\widetilde{\mathbb{P}}\text{-a.s., as } n \to \infty.$$

Thus, using $u(s) = b^*$ for $s \in S$, we obtain

$$\liminf_{n \to \infty} \frac{(\sup_{t \geqslant 0} \widetilde{U}^n(t) - nb^*)}{\sqrt{n}} \geqslant \sup_{s \in S} \widetilde{W}(s), \quad \widetilde{\mathbb{P}}\text{-a.s.} \tag{14}$$

Let τ^n be the smallest maximizer of $\widetilde{U}^n(\cdot)$, i.e., $\tau^n = \inf\{t \colon \widetilde{U}^n(t) = \sup_{s \geqslant 0} \widetilde{U}^n(s)\}$. The existence of τ^n follows from the fact that $\sup_{s \geqslant 0} \widetilde{U}^n(s) < \infty$ because $\widetilde{U}^n(0) = 0$ and $\lim_{t \to \infty} \widetilde{U}^n(t) = 0$, and the point at which the supremum is first achieved is well defined because $\widetilde{U}^n$ has piecewise constant RCLL sample paths. To see that τ^n is a measurable random variable, note that we can write $\{\tau^n \leqslant y\} = \{\sup_{t \geqslant y} \widetilde{U}^n(t) \leqslant \sup_{t \leqslant y} \widetilde{U}^n(t)\}$, and because $\widetilde{U}^n$ has RCLL paths, the quantities $\sup_{t \geqslant y} \widetilde{U}^n(t)$ and $\sup_{t \leqslant y} \widetilde{U}^n(t)$ are measurable (arguing as in the proof of Lemma 1.2).

Fix an $\omega \in \widetilde{\Omega}$ such that the convergence in (13) holds. Using the definition of τ^n, we can write

$$\begin{aligned}\frac{(\sup_{t \geqslant 0} \widetilde{U}^n(t) - nb^*)}{\sqrt{n}} &= \frac{(\widetilde{U}^n(\tau^n) - nb^*)}{\sqrt{n}} \\ &\leqslant \frac{(\widetilde{U}^n(\tau^n) - nu(\tau^n))}{\sqrt{n}},\end{aligned}$$

where the inequality follows from the fact that $u(\tau^n) \leqslant b^*$. We now establish that any cluster point of $\{\tau^n\}$ must lie in S. We proceed by a contradiction argument. Assume the contrary, i.e., there exists a cluster point of $\{\tau^n\}$ that does not lie in S. Then, there exists a subsequence $\{\tau^{n_k}\}$ such that $\tau^{n_k} \to \tau \in [0,\infty] \backslash S$ as $k \to \infty$. Then, we have

$$\begin{aligned}\left| \frac{\widetilde{U}^{n_k}(\tau^{n_k})}{n_k} - u(\tau) \right| &\leqslant \left| \frac{\widetilde{U}^{n_k}(\tau^{n_k})}{n_k} - u(\tau^{n_k}) \right| + |u(\tau^{n_k}) - u(\tau)| \\ &\leqslant \sup_{t \geqslant 0} \left| \frac{\widetilde{U}^{n_k}(t)}{n_k} - u(t) \right| + |u(\tau^{n_k}) - u(\tau)|,\end{aligned}$$

which combined with the continuity of u gives us

$$\frac{\widetilde{U}^{n_k}(\tau^{n_k})}{n_k} \to u(\tau) < \sup_{t \geqslant 0} u(t)$$

as $k \to \infty$, where the inequality is strict as $\tau \notin S$. However, we can rewrite (12) as $\widetilde{U}^n(\tau^n)/n \to \sup_{t \geqslant 0} u(t)$ as $n \to \infty$, which gives us the required contradiction. Thus, for any subsequence denoted by n_k, there exists a further subsequence denoted by n_{k_ℓ} such that $\tau^{n_{k_\ell}} \to \tau \in S$. Using the convergence in (13) and the continuity of $\widetilde{W}$, we obtain

$$\lim_{\ell \to \infty} \frac{(\widetilde{U}^{n_{k_\ell}}(\tau^{n_{k_\ell}}) - n_{k_\ell} u(\tau^{n_{k_\ell}}))}{\sqrt{n_{k_\ell}}} = \widetilde{W}(\tau) \leqslant \sup_{s \in S} \widetilde{W}(s),$$

which immediately gives us

$$\limsup_{n \to \infty} \frac{(\widetilde{U}^n(\tau^n) - nu(\tau^n))}{\sqrt{n}} \leqslant \sup_{s \in S} \widetilde{W}(s). \tag{15}$$

Combining (14) and (15), we obtain

$$\sqrt{n}\left(\sup_{t \geqslant 0} \frac{\widetilde{U}^n(t)}{n} - b^*\right) \to \sup_{s \in S} \widetilde{W}(s), \quad \widetilde{\mathbb{P}}\text{-a.s., as } n \to \infty,$$

and the result follows. □

Proof of Proposition 3. Using the characterization of b^* in (8), we obtain the asymptotic peak of the usage process corresponding to the distribution F_ϵ, $b^*_{F_\epsilon} = \max(\epsilon, m\epsilon + \epsilon^2 - \epsilon^3)$. Thus, $b^*_{F_\epsilon} = O(\epsilon)$. Further, we can compute $\int_0^\infty x^2 dF_\epsilon(x) = m^2/\epsilon - 2m + \epsilon(2m+1) - \epsilon^2$. Thus, the second moment of this distribution is $O(1/\epsilon)$. This completes the proof of part 1. Noting that ϵ is arbitrary, part 2 immediately follows. □

Proof of Proposition 4. Using (8), we have

$$\begin{aligned} b^* &= \sup_{t \geqslant 0} \frac{\int_0^t \exp(-s/m)\exp(-(t-s)/m)\,ds}{m} \\ &= \sup_{t \geqslant 0} \frac{t}{m} \exp\left(-\frac{t}{m}\right) \\ &= \sup_{t \geqslant 0} te^{-t} \\ &= e^{-1}. \quad \square \end{aligned}$$

Proof of Proposition 5. As stated in (10),

$$b^*_d = \sup_{0 \leqslant t \leqslant 2d} \sum_{\theta \in \Theta} p_\theta \frac{\int_0^t (1 - F^d_\theta(s))(1 - F^d_\theta(t-s))\,ds}{\int_0^d (1 - F^d_\theta(s))\,ds},$$

where F^d_θ is the rental time distribution of a type θ customer truncated at d. Define

$$A_\theta(t, d) \equiv \frac{\int_0^t (1 - F^d_\theta(s))(1 - F^d_\theta(t-s))\,ds}{\int_0^d (1 - F^d_\theta(s))\,ds}$$

for $\theta \in \Theta$ and $0 \leqslant t \leqslant 2d$. Noting that $F^d_\theta(s) = 1$ for $s \geqslant d$, we obtain for each $t \geqslant d$,

$$A_\theta(t, d) \equiv \frac{\int_0^d (1 - F^d_\theta(s))(1 - F^d_\theta(t-s))\,ds}{\int_0^d (1 - F^d_\theta(s))\,ds}.$$

Because $F^d_\theta(t-s)$ is increasing in t for each s, we obtain that $A_\theta(t, d)$ is nonincreasing for $t \geqslant d$, and we have

$$b^*_d = \sup_{0 \leqslant t \leqslant 2d} \sum_{\theta \in \Theta} p_\theta A_\theta(t, d) = \sup_{0 \leqslant t \leqslant d} \sum_{\theta \in \Theta} p_\theta A_\theta(t, d).$$

To prove the result, it suffices to show that for any $d_1 \leqslant d_2$ and $t \in [0, d_2]$, there exists a $\hat{t} \in [0, d_1]$ such that $A_\theta(\hat{t}, d_1) \geqslant A_\theta(t, d_2)$ for all $\theta \in \Theta$.

Case I: For $t \leqslant d_1$, we can write

$$\frac{A_\theta(t, d_1)}{A_\theta(t, d_2)} = \frac{\int_0^{d_2} (1 - F_\theta(s))\,ds}{\int_0^{d_1} (1 - F_\theta(s))\,ds}.$$

Thus, noting that $\int_0^d [1 - F^d_\theta(s)]\,ds$ is increasing in d, we obtain $A_\theta(t, d_1) \geqslant A_\theta(t, d_2)$ for all $\theta \in \Theta$.

Case II: We consider $t \in (d_1, d_2]$. We show that $A_\theta(d, d)$ is nonincreasing in d. Note that $F^d_\theta(s) = F_\theta(s)$ for $s < d$, and thus we can write

$$A_\theta(d, d) = \frac{\int_0^d (1 - F_\theta(s))(1 - F_\theta(d-s))\,ds}{\int_0^d (1 - F_\theta(s))\,ds}.$$

The derivative of $A_\theta(d, d)$ with respect to d is given by

$$\begin{aligned} &\frac{\int_0^d (1 - F_\theta(s))\,ds[(1 - F_\theta(d)) - \int_0^d (1 - F_\theta(s))\,dF_\theta(d-s)]}{(\int_0^d (1 - F_\theta(s))\,ds)^2} \\ &\quad - \frac{(1 - F_\theta(d))\int_0^d (1 - F_\theta(s))(1 - F_\theta(d-s))\,ds}{(\int_0^d (1 - F_\theta(s))\,ds)^2} \\ &\overset{(a)}{\leqslant} \left(\left[\int_0^d (1 - F_\theta(s))\,ds\right](1 - F_\theta(d))\left[1 - \int_0^d dF_\theta(d-s)\right]\right. \\ &\qquad \left. - (1 - F_\theta(d))\int_0^d (1 - F_\theta(s))(1 - F_\theta(d-s))\,ds\right) \\ &\quad \cdot \left(\int_0^d (1 - F_\theta(s))\,ds\right)^{-2} \\ &= \frac{1 - F_\theta(d)}{(\int_0^d (1 - F_\theta(s))\,ds)^2}\left[\int_0^d (1 - F_\theta(s))(F_\theta(d-s) - F_\theta(d))\,ds\right] \\ &\overset{(b)}{\leqslant} 0, \end{aligned}$$

where (a) follows as $F_\theta(s) \leqslant F_\theta(d)$ and (b) follows by the monotonicity of $F_\theta(\cdot)$. Thus, we have $A_\theta(t, t) \leqslant A_\theta(d_1, d_1)$ for all $\theta \in \Theta$. Repeating the argument in Case I, we have $A_\theta(t, t) \geqslant A_\theta(t, d_2)$ for all $\theta \in \Theta$. Combining these inequalities, we obtain $A_\theta(d_1, d_1) \geqslant A_\theta(t, d_2)$, and this completes the proof. □

Proof of Proposition 6. The first part of the result follows from Corollary 1.2. The second part of the proof follows by noting that the stockout probability is monotone in b^n. □

Proof of Proposition 7. Pick a sample path $\omega \in \Omega$ on which $\sup_{t \geqslant 0} U^n(t)/n \to b^*$ as in Corollary 1.1. Then,

there exists $N_\epsilon(\omega) < \infty$ such that on ω, $\sup_{t\geqslant 0} U^n(t) \leqslant n(b^* + \epsilon) = b^n$ for $n > N_\epsilon(\omega)$. Noting that if $\sup_{t\geqslant 0} U^n(t) \leqslant b^n$, then $d^n(b^n) = 0$, and the result follows. □

References

Adler, R. J., J. Taylor. 2007. Random fields and geometry. *Springer Monographs in Mathematics*. Springer, New York.

Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* **57** 66–81.

Bassamboo, A., R. S. Randhawa. 2009. Optimal control in a Netflix-like closed rental system. Working paper, Northwestern University, Evanston, IL.

Billingsley, P. 1995. *Probability and Measure*. John Wiley and Sons, New York.

Chen, H., D. D. Yao. 2001. *Fundamentals of Queueing Networks*. Springer-Verlag, New York.

Ethier, S. N., T. G. Kurtz. 1986. *Markov Processes*: *Characterization and Convergence*. John Wiley and Sons, New York.

Glynn, P. W., W. Whitt. 1991. A new view of the heavy-traffic limit theorem for many-server queues. *Adv. Appl. Probab.* **23** 188–209.

Gromoll, H. C., A. L. Puha, R. J. Williams. 2002. The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.* **12** 797–859.

Gromoll, H. C., P. Robert, B. Zwart. 2008. Fluid limits for processor sharing queues with impatience. *Math. Oper. Res.* **33** 375–402.

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

Kiesmüller, G., E. van der Laan. 2001. An inventory model with dependent product demands and returns. *Internat. J. Production Econom.* **72** 73–87.

Krichagina, E. V., A. A. Puhalskii. 1997. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems*: *Theory Appl.* **25** 235–280.

Mortimer, J. 2008. Vertical contracts in the video rental industry. *Rev. Econom. Stud.* **75** 165–199.

Puhalskii, A., M. I. Reiman. 2000. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **32** 564–595.

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

Reed, J. E. 2007. The G/GI/N queue in the Halfin-Whitt regime I: Infinite server queue system equations. Working paper, New York University.

Tang, C. S., S. Deo. 2008. Rental price and rental duration under retail competition. *Eur. J. Oper. Res.* **187** 806–828.

van der Vaart, A. W. 2000. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

Whitt, W. 2005. Heavy-traffic limits for the G/H2/N/M queue. *Math. Oper. Res.* **30** 1–27.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, NJ.