Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Sensitivity of Optimal Capacity to Customer Impatience in an Unobservable M/M/S Queue (Why You Shouldn't Shout at the DMV)

Mor Armony, Erica Plambeck, Sridhar Seshadri,

To cite this article:
Mor Armony, Erica Plambeck, Sridhar Seshadri, (2009) Sensitivity of Optimal Capacity to Customer Impatience in an Unobservable M/M/S Queue (Why You Shouldn't Shout at the DMV). Manufacturing & Service Operations Management 11(1):19-32. http://dx.doi.org/10.1287/msom.1070.0194

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 11, No. 1, Winter 2009, pp. 19–32
ISSN 1523-4614 | EISSN 1526-5498 | 09 | 1101 | 0019

informs

DOI 10.1287/msom.1070.0194

Sensitivity of Optimal Capacity to Customer Impatience in an Unobservable M/M/S Queue (Why You Shouldn't Shout at the DMV)

Mor Armony
Stern School of Business, New York University, New York, New York 10012, marmony@stern.nyu.edu

Erica Plambeck
Graduate School of Business, Stanford University, Stanford, California 94305, elp@stanford.edu

Sridhar Seshadri
Stern School of Business, New York University, New York, New York 10012, sseshadr@stern.nyu.edu

This paper employs sample path arguments to derive the following convexity properties and comparative statics for an *M/M/S* queue with impatient customers. If the rate at which customers balk and renege is an increasing, concave function of the number of customers in the system (head count), then the head-count process and the expected rate of lost sales are decreasing and convex in the capacity (service rate or number of servers). This result applies when customers cannot observe the head count, so that the balking probability is zero and the reneging rate increases linearly with the head count. Then the optimal capacity increases with the customer arrival rate but is not monotonic in the reneging rate per customer. When capacity is expensive or the reneging rate is high, the optimal capacity decreases with any further increase in the reneging rate. Therefore, managers must understand customers' impatience to avoid building too much capacity, but customers have an incentive to conceal their impatience, to avoid a degradation in service quality. If the system manager can prevent customers from reneging during service (by requiring advance payment or training employees to establish rapport with customers), the system's convexity properties are qualitatively different, but its comparative statics remain the same. Most important, the prevention of reneging during service can substantially reduce the total expected cost of lost sales and capacity. It increases the optimal capacity (service rate or number of servers) when capacity is expensive and reduces the optimal capacity when capacity is cheap.

Key words: capacity planning; queueing systems; reneging; balking; unobservable queues; stochastic convexity; sample path convexity
History: Received: June 8, 2005; accepted: June 28, 2007. Published online in *Articles in Advance* January 4, 2008.

1. Introduction

This paper develops qualitative insights about how the optimal capacity investment for a make-to-order manufacturing or service system is influenced by customers' impatience, which may lead them to cancel an order (renege) or to not order at all (balk) when waiting is required. Technically, we prove convexity and comparative statics properties for *M/M/S* queues with state-dependent reneging and balking.

A dominant assumption in the manufacturing operations management literature is that customers will wait for as long as necessary to obtain a product (infinite backordering). In reality, only a subset of customers will wait and only for a limited time. Unfortunately, models that incorporate dynamic balking and reneging are notoriously intractable. There exist many structural results and simple optimal policies for inventory management with infinite backordering, but relatively few exist for systems with lost sales, and these require strong assumptions (e.g., at most one order may be outstanding (Johansen and Thorstenson 1993, 1996; Moinzadeh and Nahmias 1988)) or approximations (Nahmias 1979, Cohen et al. 1988, Johansen and Hill 2000). A number of papers provide analytic results for make-to-order manufacturing systems in which the customer arrival process depends on the static expected waiting cost, but not

dynamic state information (Mendelson and Whang 1990; Van Mieghem 1995, 2000; Armony and Haviv 2003; Lederer and Li 1997; Afeche 2004). Make-to-order systems with state-dependent lead time quotation, which drives state-dependent balking, are much more complex, so researchers employ heuristic algorithms, simulation, and approximations (Duenyas and Hopp 1995, Hopp and Sturgis 2001, Keskinocak et al. 2001, Kapuscinski and Tayur 2007, Plambeck 2004). All these papers model the make-to-order system with a single-server queue. In contrast, we provide analytic results for multiserver systems with state-dependent balking and reneging.

Modeling balking and reneging is difficult but worthwhile, because one obtains new managerial insights. Ho et al. (2002) and Kumar and Swaminathan (2003) incorporate reneging and balking, respectively, into the well-known Bass model of new product introduction and obtain qualitatively different insights and structurally different optimal control policies. Failure to account for duplicate ordering and reneging can cause either over- or underinvestment in manufacturing capacity (Armony and Plambeck 2005). For certain assemble-to-order systems in which a customer reneges after a deterministic amount of time, independent control of each component is optimal (Plambeck and Ward 2007, Plambeck 2004).

In contrast, when customers wait for as long as necessary to obtain the product, optimal production for each component depends on the inventory positions of all the other components (Plambeck and Ward 2006). In Li and Lee (1994) two firms compete by setting prices; customers observe queue lengths and jockey between the firms to minimize delivery time. In contrast to traditional Bertrand equilibrium with zero prices and profits, because customer orders depend dynamically on the lead time, the firms sustain strictly positive profits. In a dynamic Bayesian formulation, Chen and Plambeck (2008) show the value of reducing inventory levels to learn about customers' balking behavior.

Most of the existing research assumes the simplest structure for reneging (customers renege after an exponentially distributed amount of time) or balking (customers balk with probability p if there is any wait and with probability $(1-p)$ until the product is delivered). Four notable exceptions are Ward and Glynn (2005), Reed and Ward (2008), Zeltyn and Mandelbaum (2005), and Whitt (2006a). These four innovative papers allow general distributions for reneging and balking and perform asymptotic analysis of these systems under conventional heavy traffic or the many-servers heavy traffic regimes. Mandelbaum and Shimkin (2000) derive complex dynamic customer behavior from primitives on valuation and waiting costs for an *M/M/S* queue with congestion/failure shocks, assuming customers cannot observe the queue length.

Most of the literature on balking and reneging in queues focuses on performance evaluation and estimation (see, for example, Baccelli and Hebuterne 1981, Garnett et al. 2002, Mandelbaum and Zeltyn 1998, and Brown et al. 2005 and references therein). Recently, however, researchers have employed diffusion approximations to characterize the optimal number of servers for large-scale call centers (Harrison and Zeevi 2005, Mandelbaum and Zeltyn 2006, Whitt 2006b) and a near-optimal admission control policy for a queueing system with reneging (Ward and Kumar 2006).

Exact comparative statics are notoriously difficult to establish by direct manipulation of expected cost functions, even for simple queueing systems with closed-form expressions for system performance (Shanthikumar and Yao 1989). We adopt the sample path approach of Shaked and Shanthikumar (1988) to establish convexity and comparative statics properties for *M/M/S* queueing systems in which customers are impatient and cannot observe the head count. Specifically, in §2 we formulate our model (an *M/M/S* queue with balking and reneging). In §3, we establish the convexity of the head-count process and related cost functions, with respect to capacity (service rate or number of servers). In §4, we assume that customers cannot observe the head count and that the total reneging rate is linear in either the head count or the queue length and investigate how customers' reneging behavior affects the optimal capacity and cost. In §5, we summarize the resulting managerial insights. All proofs are in the appendix.

2. Model Formulation

Either a make-to-order manufacturing system or a service system may be modeled by a multiserver,

infinite-buffer queueing system as described here. Customers arrive at the system according to a Poisson process with rate λ. The service time has an exponential distribution with rate μ, and the system has S servers. We denote the number of customers in the system (including those in service) at time t by $Y(t)$ and refer to that number as the head count. Let y be a generic realization of the head count (system state) at an arbitrary time point. An arriving customer may decide to balk, namely, to leave on arrival. The balking probability is a function of the head count and is denoted by $\beta(y)$. Finally, customers may decide to cancel their order (renege) at any point during their wait or while being served. The reneging rate is a function of the head count and is denoted by $\eta(y)$. All arrivals, service times, balking, and reneging are assumed to be independent. Therefore, the head-count process $Y = \{Y(t), t \geq 0\}$ is a continuous time Markov chain.

The system manager knows the customers' characteristics, modeled here by λ, β, and η, and wishes to choose μ, the "capacity per server," to minimize the cost associated with lost sales, holding customers in the system, and capacity investment:

$$C(\mu, S; \lambda, \eta, \beta) = c[\lambda E\beta(Y(\infty)) + E\eta(Y(\infty))] + Eh(Y(\infty)) + k\mu S, \quad (1)$$

where $Y(\infty)$ is the head count in steady state; without loss of generality, it is assumed that $c = 1$. Make-to-order manufacturing is commonly modeled by a single-server queue, in which capacity investment naturally corresponds to the choice of service rate μ. Service systems (e.g., a call center with S operators) are commonly modeled as multiserver queues. In service operations, the choice of μ could represent an investment in training or support software to make each operator more productive. However, the service system manager is more likely to be concerned with choosing the number of servers S than the capacity per server μ. Therefore, we will develop parallel results regarding the choice of S to minimize the right-hand side of (1). In make-to-order manufacturing, the holding cost $h(\cdot)$, includes loss of goodwill and includes clerical effort because some customers modify their orders while waiting. It may also include late penalty fees specified by a contract. In service operations, the holding cost also includes loss of goodwill, and in some cases it includes the actual cost of physically holding space (such as parking spots in restaurants and trunk lines in call centers).

We will employ the following technical definitions. Following Shaked and Shanthikumar (1988), we say that a process Y is stochastically decreasing and convex in a parameter μ in the sample path sense (SDCX(sp)) if, for any $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ such that $\mu_1 + \mu_4 = \mu_2 + \mu_3$, there exist Y_1, Y_2, Y_3, and Y_4 that are versions of the original process Y corresponding to $\mu = \mu_1, \mu_2, \mu_3$, and μ_4, respectively, and that satisfy the following two properties for all $t \geq 0$:

1. $Y_1(t) + Y_4(t) \geq Y_2(t) + Y_3(t)$, a.s., and
2. $Y_1(t) \geq \max\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s. (where a.s. stands for almost surely).

Similarly, Y is stochastically increasing and concave in a parameter μ in the sample path sense (SICV(sp)) if, for any $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ such that $\mu_1 + \mu_4 = \mu_2 + \mu_3$, there exist Y_1, Y_2, Y_3, and Y_4 that are versions of the original process Y corresponding to $\mu = \mu_1, \mu_2, \mu_3$, and μ_4, respectively, and that satisfy the following two properties for all $t \geq 0$:

1. $Y_1(t) + Y_4(t) \leq Y_2(t) + Y_3(t)$, a.s., and
2. $Y_1(t) \leq \min\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s.

Finally, let θ be an arbitrary parameter, and let $\mu^*(\theta)$ be a value of μ that minimizes a certain function $g(\mu; \theta)$. The meaning of the saying "$\mu^*(\theta)$ is increasing in θ" when $\mu^*(\theta)$ is not necessarily unique is that the set of minimizers $\mu^*(\theta)$ of $g(\mu; \theta)$ is *ascending* in θ. (See, for example, Porteus 2002.) Throughout the paper we use the term *increasing* to mean nondecreasing and the term *decreasing* to mean nonincreasing.

3. Convexity Properties of Systems with Balking and Reneging

Our first result is that the head-count process and related cost functions are convex in the service rate μ and number of servers S. This result is very general, requiring only that the reneging rate and balking probability are increasing and concave in the head count.

THEOREM 1. *Suppose that the reneging rate $\eta(\cdot)$ and the balking probability $\beta(\cdot)$ are both increasing and concave functions of the head count y. The head-count process Y is SDCX(sp) in the capacity per server, μ, and is also SDCX(sp) in the number of servers, S. Furthermore, if $h(\cdot)$*

is increasing and convex, then $Eh(Y(\infty))$ is decreasing and convex in μ and in S.

Shaked and Shanthikumar (1988) established a result similar to Theorem 1: In an exponential single-server queue with a departure rate that is increasing and concave in the head count, the head-count process is stochastically increasing and convex (SDCX(sp)) in the arrival rate.

Our assumption in Theorem 1 that the reneging rate is increasing and concave in the head count Y is natural, especially when customers cannot observe the head count. The assumption that the balking probability is increasing and concave in the head count Y is natural in the case of a single server but may be violated in systems with multiple servers. We will demonstrate these points with two fundamental models of heterogeneous customers' preferences.

The first preference model is most plausible when customers cannot observe the head count. The ith customer arrives in the system at time τ_i and will wait for at most T_i time units. That is, he reneges from the queue at time $\tau_i + T_i$ if he has not yet completed service. The balking probability is zero. The T_i are independent and identically distributed, with expected value $E[T]$ and hazard rate $\hat{h}(\cdot)$. Of course, if the T_i are exponentially distributed, then the reneging rate is linear in the head-count process: $\eta(y) = y/E[T]$. For generally distributed T_i, Whitt (2005) shows that the $M/M/S+GI$ system can be effectively approximated by our Markovian system with state-dependent reneging rate $\tilde{\eta}(y)$. In Whitt's approximation, $\tilde{\eta}(y)$ is set equal to the intensity of reneging for the $M/M/S+GI$ in steady state, conditional on $Y(t) = y$. Recently, other researchers have characterized the shape of $\tilde{\eta}(y)$. Brandt and Brandt (2002) prove that the intensity of reneging $\tilde{\eta}(y)$ is nondecreasing in y and that $\tilde{\eta}(y)$ is asymptotically linear as the head count grows large:

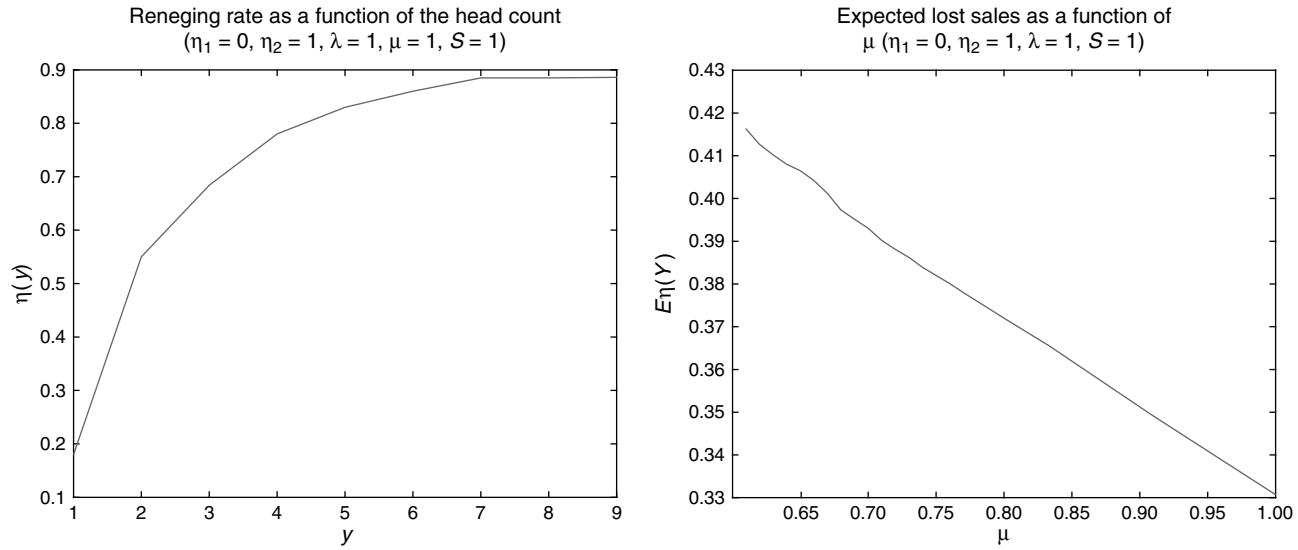
$$\lim_{y\to\infty}(\tilde{\eta}(y+1) - \tilde{\eta}(y)) = 1/E[T].$$

Reed and Ward (2008) characterize the limiting distribution of the $GI/GI/1+GI$ queue in heavy traffic.[1] As an immediate corollary to their Theorem 1, *the intensity of reneging $\tilde{\eta}(y)$ is concave in the head count if and only if $\hat{h}(\cdot)$ is a decreasing function*. Although Reed and Ward (2008) assume a single server, the result can be extended to a finite number of servers. The logic underlying this result is that the customer in the ith position from the end of queue has been waiting for approximately time i/λ. Therefore, the conditional probability that this customer reneges in the next δ units of time is approximately $\delta\hat{h}(i)\lambda$, so the total reneging rate is approximately $\sum_{i=1}^{Y}\hat{h}(i/\lambda)$, which is concave in Y if and only if $\hat{h}(\cdot)$ is a decreasing function. In most practical applications, customers are heterogeneous in their patience, so a decreasing hazard rate is the natural modeling assumption. To see this, suppose that an arriving customer is equally likely to be "patient" or "impatient." If she is patient, her maximum wait T_i is exponentially distributed with rate η_1. If she is impatient, her maximum wait T_i is exponentially distributed with rate η_2 where $\eta_1 < \eta_2$. A customer who has waited for some time without reneging is more likely to be of the "patient" type than a new arrival. The rate of reneging by customers at the end of the queue is higher than by customers at the head of the queue. Although an exponential random variable has a constant hazard rate, any such mixture of exponential random variables has a decreasing hazard rate. Therefore, by Theorem 1 in Reed and Ward (2008), the intensity of reneging $\tilde{\eta}(y)$ will be approximately concave in the head count. The left panel in Figure 1 shows the intensity of reneging as a function of the head count for a single-server system with $\eta_1 = 0$, $\eta_2 = 1$, $\mu = 1$, and $\lambda = 1$. The right panel of Figure 1 shows the expected reneging rate as a function of the capacity μ.

The second preference model is relevant when customers observe the head count. The ith customer to arrive has valuation for service V_i and cost of waiting c per unit time. The V_i are independent and identically distributed. The price of service (or of the product in a make-to-order system) is fixed at p. Then customers may balk but will not renege. In the

[1] In their heavy traffic scaling, the arrival rate λ is of order n, the capacity imbalance $(\mu - \lambda)/\mu$ is of order $\sqrt{n}$, and the maximum waiting times T_i are of order $1/\sqrt{n}$ and $n \to \infty$. Related papers prove that this heavy traffic scaling of the arrival rate, capacity imbalance, and waiting times arises naturally from optimal capacity investment and/or pricing decisions (Borst et al. 2004; Maglaras and Zeevi 2003; Plambeck and Ward 2006, 2007; Plambeck 2004; Randhawa and Kumar 2005).

Figure 1 Reneging Rate and Expected Lost Sales in the Heterogeneous Customer Model



single-server case, the ith arriving customer, observing y customers ahead of her, balks if and only if $V_i < p + cy/\mu$. Therefore, the balking probability is concave if and only if the price p is set such that the cumulative distribution function $\Pr(V_i \le v)$ is concave for $v \ge p$. This will always be true for uniformly and exponentially distributed valuation V_i and is true for normally distributed valuation if and only if $p \ge E[V_i]$. With multiple ($S \ge 2$) servers, the balking probability is constant for $y = 0, 1, \ldots, S-1$ and strictly increasing with y for $y \ge S$ and therefore cannot be concave.

For some strictly concave balking and reneging functions, the cost of lost sales $c[\lambda E\beta(Y(\infty)) + E\eta(Y(\infty))]$ is not convex in capacity. For example, in the right panel of Figure 1, the expected reneging rate is strictly concave in μ (although almost imperceptibly so) for $\mu \in [0.65, 0.69]$. Therefore, the cost function (1) will be convex only if the holding cost $h(\cdot)$ is sufficiently convex.

4. Sensitivity of Optimal Capacity and Cost to Customers' Reneging Behavior

In many manufacturing and service systems, customers cannot see the head count, and the costs of capacity and lost sales dominate any holding costs. (For example, on a recent visit to the Department of Motor Vehicles [DMV], one of the authors was assigned a letter, indicating the type of transaction, and a number. Although the author saw multitudes of people sitting or milling about the lobby and saw various letter-number combinations flashing above the 20-odd servers to indicate which customer would be served next, she could not ascertain her waiting time or position in line. The head-count process was effectively unobservable, despite the author's best effort. The DMV was patently unconcerned about the author's waiting cost, which translates into $h(\cdot) = 0$ in our model.) Throughout this section, we assume that $h(\cdot) = 0$ and customers cannot observe the head count, which implies that they do not balk: $\beta(\cdot) = 0$. The purpose of this section is to investigate how customers' reneging behavior affects the optimal capacity and cost. To perform this sensitivity analysis, we must begin by characterizing the convexity properties of our cost function.

First, we consider the case that the reneging rate is linear in the head count: $\eta(y) = \eta y$ for some positive constant η. Equivalently, each customer will renege after an exponentially distributed amount of time with rate η. As discussed in the previous section, this is a reasonable modelling assumption in systems where customers cannot observe the head count. By

application of Theorem 1, we find that the cost function is convex in the service rate μ and in the number of servers S.

PROPOSITION 1. *Suppose that the reneging rate* $\eta(y) = \eta y$. *Then the cost function*

$$C(\mu, S; \lambda, \eta) = \eta E Y(\infty) + k\mu S \tag{2}$$

is convex in μ *and in* S.

Convexity of the cost function allows us to evaluate how changes in the model parameters affect the optimal capacity. Theorem 2 establishes that the optimal capacity increases with the customer arrival rate λ, as one might expect. Surprisingly, as η increases, meaning that customers become more impatient, the optimal capacity may decrease. For a fixed S, $\mu^*(\lambda, \eta)$ denotes the optimal service rate that minimizes (2), and, for a fixed μ, $S^*(\lambda, \eta)$ denotes the optimal number of servers that minimizes (2).

THEOREM 2. *Suppose that the reneging rate* $\eta(y) = \eta y$. *Both the optimal service rate* $\mu^*(\lambda, \eta)$ *and the optimal number of servers* $S^*(\lambda, \eta)$ *increase with* λ *but can either increase or decrease with* η.

The proof of Theorem 2 employs Proposition 1 and the notion of sample-path submodularity.

Second, we consider the difficult-to-analyze case that the reneging rate is linear in the queue length: $\eta(y) = \eta(y - S)^+$ for some positive constant η. Equivalently, each customer will renege after an exponential time with rate η as long as she is waiting in line but will not renege during service. This "commitment during service" is a plausible assumption when customers cannot observe the head count but do experience the service; i.e., a customer knows when his service is in progress. In particular, the proposed model of reneging is applicable to call centers or the DMV. With commitment during service, the reneging rate $\eta(y)$ is *not* concave in the head count y, so Theorem 1 is not applicable. Nevertheless, we are able to establish some convexity properties of the cost function with respect to the capacity decision variables μ and S. We can then use those convexity properties in numerically evaluating the sensitivity of the optimal capacity and cost to customers' reneging behavior.

The analog of Theorem 1 for systems with customer commitment is relatively complex.

THEOREM 3. *Suppose that the reneging rate* $\eta(y) = \eta(y - S)^+$. *If* $\mu < \eta$, *then the head-count process* Y *is stochastically increasing and concave* (*SICV*(*sp*)) *in* S *but stochastically decreasing in* μ. *If* $\mu \geq \eta$, *then* Y *is stochastically decreasing and convex* (*SDCX*(*sp*)) *in* S *and in* μ.

Our sample-path arguments fail to establish convexity or concavity of the head-count process in μ for $\mu < \eta$. However, an extensive numerical study suggests that $EY(\infty)$ is convex in μ for all $\mu \geq 0$. The intuition behind Theorem 3 is that each server plays a dual role—to grab and hold a customer out of the queue and prevent her reneging. This dual role is most important when the reneging rate for a customer waiting in the queue, η, is large. When η is large, the head count stochastically *increases* with the number of servers S. In contrast, when η is relatively small, the head count stochastically decreases with S.[2] By comparing Theorem 3 with Theorem 1, we conclude that customer commitment during service causes the head count to be stochastically *concave* in the number of servers S (rather than convex) when η is large. This is intuitive, as we expect that as S increases, the marginal absolute difference in the head-count resulting from adding another server would decrease. When Y is decreasing in S, this implies convexity, which is consistent with the noncommitment case. However, when $\mu \leq \eta$, Y is increasing in S, and in this case our intuition is consistent with concavity.

By application of Theorem 3, we find that, in the parameter region $\mu \geq \eta$, our cost function is convex in the service rate μ and in the number of servers S.

PROPOSITION 2. *Suppose that the reneging rate* $\eta(y) = \eta(y - S)^+$. *Then in the parameter region* $\mu \geq \eta$, *the cost function*

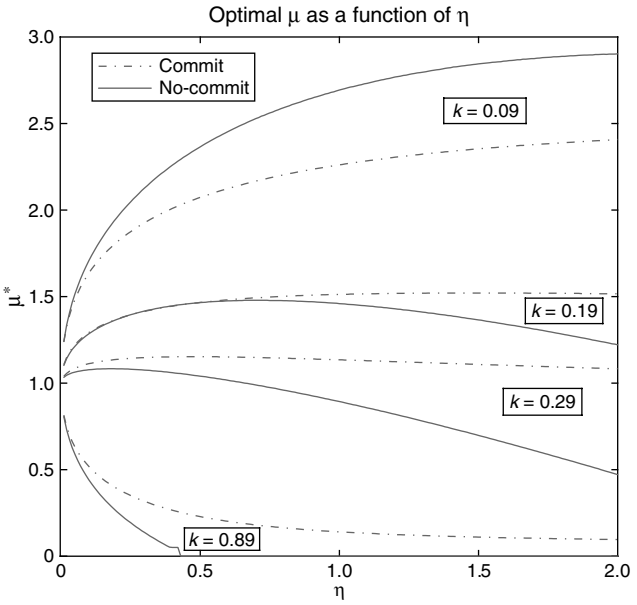
$$C(\mu, S; \lambda, \eta) = \eta E[Y(\infty) - S]^+ + k\mu S \tag{3}$$

is convex in μ *and in* S.

REMARK 1. Koole and Pot (2006) incorporate buffer capacity (maximum queue length) as a decision variable, along with number of servers S, and minimize the cost in (3) plus a linear buffer cost. In contrast to Proposition 2, they show that the cost function is not

[2] This stochastic monotonicity property has been established in Feldman et al. (2008).

Figure 2 **Optimal Capacity μ^* as a Function of the Reneging Rate η ($S = 1$, $\lambda = 1$)**



necessarily convex. (They give an example of a local minimum that is not a global minimum.) This phenomenon occurs in their model because of the discrete nature of the two decision variables. (More details are provided in the technical appendix, located in the online companion to this paper).

We are now prepared to investigate how reneging behavior affects the optimal capacity and cost. The three figures in this section provide qualitative insights that are representative of the results we obtained in an extensive numerical study.

Figure 2 shows the optimal capacity μ for a single-server queue as a function of the reneging rate per customer η and shows whether customers are committed (not to renege) during service. First consider the impact of increasing η. When the cost per unit capacity k is sufficiently low, the optimal capacity μ strictly increases with η. (Intuitively, as customers become more impatient, the system manager must build more capacity to avoid losing more customers.) In contrast, when capacity is very expensive, the optimal capacity μ strictly decreases with η. To develop intuition for that result and the nonmonotonicity of the optimal capacity μ in η for moderate levels of capacity cost k, observe that as η increases, the amount of capacity μ required to achieve a target rate of service completion ($\lambda - E\eta(Y(\infty))$) also increases. Equivalently, as η increases, the ratio $k\mu/(\lambda - E\eta(Y(\infty)))$ increases, which means that the *effective* cost of capacity per service completion increases. Therefore, for any fixed k, for sufficiently large η, as customers become even more impatient, the system manager should simply resolve to lose more customers rather than build more capacity—the optimal μ strictly decreases with η. For systems without commitment during service, for any $k > 0$,

$$\mu^*(\eta) \to 0 \quad \text{as } \eta \to \infty$$

(see Proposition 5 and its proof in the appendix). This is true even for the case $k = 0.09$ in Figure 2, where $\mu^*(\eta)$ appears strictly increasing—for sufficiently large $\eta > 2$, $\mu^*(\eta)$ decreases to 0. For systems with commitment and $k < 1$

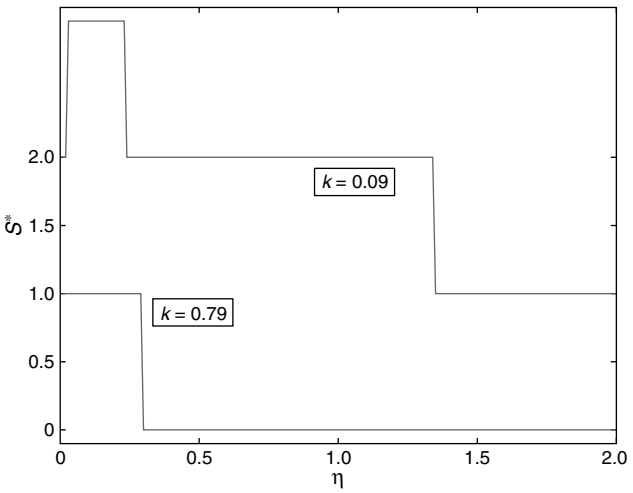
$$\mu^*(\eta) \to \lambda(k^{-1/2} - 1) > 0 \quad \text{as } \eta \to \infty,$$

which is straightforward to prove. (With or without commitment, because the penalty for a lost sale is normalized to 1, if $k \geq 1$, then the optimal capacity is zero.) The effect of customer commitment during service on the optimal capacity μ is similar to that of a reduction in η. (This is to be expected because commitment means $\eta = 0$ for customers in service.) When the cost per unit capacity k is low, the optimal capacity μ strictly decreases with commitment. Conversely, when the cost per unit capacity k is high, the optimal capacity μ strictly increases with commitment.

Figure 3 shows that the optimal number of servers varies with η qualitatively in the same manner as does the optimal capacity per server. When servers are very cheap and the reneging rate is very low, $S^*(\eta)$ initially increases with η. When servers are expensive and/or the reneging rate is large, $S^*(\eta)$ decreases with η. In our numerical experiments we have observed that this behavior is typical in both the commitment and the noncommitment cases.

From the perspective of customers of the DMV, the take-away from Figure 2 and Figure 3 is "Don't shout at the DMV!" For the DMV, the cost of capacity is very large compared with the penalty for a lost customer. (An employee at the DMV told one of the authors, "When customers get really upset [about waiting] I encourage them to do their business online.") This

Figure 3 **Optimal Capacity S^* as a Function of the Reneging Rate η ($\lambda = 1, \mu = 0.54$, Noncommitment)**



Note. Optimal S can be nonmonotone in ($\mu = 0.54$, no-commitment).

is precisely the parameter region in which an increase in the reneging rate η causes a decrease in the optimal capacity. To the extent that customers shout and threaten to leave, DMV employees may increasingly believe that customers will rapidly renege (exit and do their business online). Under optimal system management, a perceived increase in the reneging rate results in *less* service capacity and thus *worsens* the quality of service for DMV customers.

Figure 4 **Optimal Cost as a Function of the Reneging Rate η ($\lambda = 1, S = 1$)**

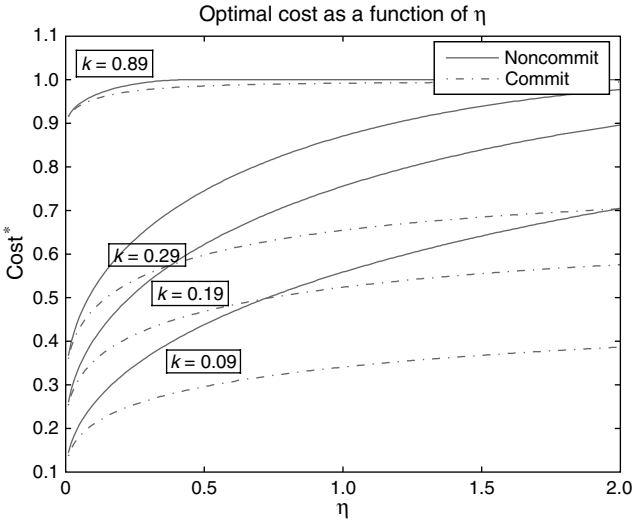


Figure 4 shows the optimal cost corresponding to the optimal capacity investment in Figure 2. As expected, cost increases with the reneging rate η. The important managerial insight here is that commitment during service greatly reduces cost and is most beneficial when the reneging rate is large. This should encourage manufacturers to disallow the cancellation of an order in process. According to Fairlie (2004), small manufacturers of customized computers charge a customer's credit card before initiating assembly, to prevent cancelations during the assembly process. Similarly, in service operations, employees should be trained to establish a rapport with customers that prevents reneging during service.

5. Conclusions

Theorem 1 establishes that the head-count process and related cost functions are decreasing and convex in the service rate and number of servers, under the assumption that the reneging rate and balking probability are increasing and concave in the head count. That assumption is most plausible for systems in which customers cannot observe the head count. In particular, when each customer reneges after an exponentially distributed amount of time with rate η, the reneging rate is linear in the head count, so Theorem 1 implies that the steady-state expected rate of lost sales is convex in the capacity (service rate or number of servers). Hence, our cost function is convex in capacity, which sets the stage for comparative statics.

The optimal capacity increases with the customer arrival rate, as one might expect, but it is not monotone in the reneging rate η. Surprisingly, when customers are impatient or capacity is expensive (relative to the cost of a lost sale), the optimal capacity decreases as customers become increasingly impatient; that is, the optimal capacity decreases with η. Therefore, when the reneging rate is high or capacity is expensive, manufacturing and service operations managers must carefully account for customers' impatience to avoid excess capacity investment. Customers, however, should *not* reveal their impatience, which would lead to a reduction in capacity investment and service quality.

We recommend that operations managers seek to prevent customers from reneging during service. This

might be accomplished by requiring payment before processing or by training service operators to establish a rapport with customers. Commitment during service destroys the concavity of the reneging rate in the head count, so Theorem 1 does not apply. When the reneging rate (for a customer in the queue) is small, the head-count process is stochastically decreasing and convex in the service rate and number of servers, as in Theorem 1. This implies that the steady-state rate of lost sales, and hence our cost function, is convex in the service rate and number of servers. However, when the reneging rate (for a customer in the queue) is large, the head-count process is stochastically increasing and concave in the number of servers, which is qualitatively opposite to the result in Theorem 1. Nevertheless, we observe that the effect of increasing the reneging rate on the optimal capacity (service rate or number of servers) is qualitatively the same in systems with and without commitment during service. Commitment during service reduces the optimal capacity when capacity is cheap (compared to a lost sale) and increases the optimal capacity when capacity is expensive. Most important, with the optimal capacity investment, commitment during service substantially reduces the steady-state expected cost of lost sales and capacity.

Electronic Companion

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

Acknowledgments

The authors are grateful to an anonymous review team that guided the development of this research, and to Maxim Afanasyev for the simulation results. This research was supported by the National Science Foundation under Grant DMI-0239840.

Appendix

Proof of Theorem 1. First consider the decision variable μ. We first prove the result for the single-server case ($S = 1$) and then deal with the general multiserver case. The proof is based on the sample path approach. Specifically, we prove that Y (viewed as a function of μ) satisfies *sample-path convexity* (a term that has been introduced by Shaked and Shanthikumar 1988). Specifically, let $0 \le \mu_1 \le \mu_2 \le \mu_3 \le \mu_4$ be four service rates such that $\mu_1 + \mu_4 = \mu_2 + \mu_3$ and fix λ, $\beta(\cdot)$, and $\eta(\cdot)$. Suppose that there exist $Y_1, \ldots, Y_4$, which are versions of the original head-count processes (Y_i has service rate of μ_i) that satisfy the following two properties for all $t \ge 0$:

1. $Y_1(t) + Y_4(t) \ge Y_2(t) + Y_3(t)$, a.s. and
2. $Y_1(t) \ge \max\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s.

Then, according to Shaked and Shanthikumar (1988), Y is said to be stochastically decreasing and convex in the sample-path sense (SDCX(sp)). From Theorem 3.6, Proposition 2.11 and Remark 2.8 of Shaked and Shanthikumar (1988), it follows that $Eh(Y(\infty))$ is decreasing and convex in μ, for any increasing and convex function h. In particular, $EY(\infty)$ is decreasing and convex in μ.

To construct the coupled versions $Y_1, \ldots, Y_4$ we wish to come up with appropriate uniformized discrete versions of the original processes. However, for uniformization to work, one needs bounded transition rates of the original Markov chain, which is not the case in this paper (we do *not* assume boundedness of the reneging rates $\eta(y)$). To resolve this problem we define for all $M > 0$ a truncated reneging function $\eta_M(y) = \min\{\eta(y), M\}$. Clearly, because $\eta(\cdot)$ is concave, and $\min\{\cdot, M\}$ is nondecreasing and concave, $\eta_M(\cdot)$ is also concave. Moreover, for any fixed $M > 0$, $\eta_M(\cdot)$ is bounded. Let $Y_1^M, \ldots, Y_4^M$ be uniformized discrete versions of the head-count processes with arrival rate λ, balking probability function $\beta(\cdot)$, service capacity μ_i, $i = 1, \ldots, 4$, and reneging rate function $\eta_M(\cdot)$. We will show that for each $M > 0$, if 1 and 2 are satisfied at time $n = 0$ with respect to $Y_1^M, \ldots, Y_4^M$, then they hold at time n for all $n \in \mathbb{Z}_+$. It will then follow that $Eh(Y^M(\infty))$ is decreasing and convex in μ. But because $Y^M(\infty)$ weakly converges to $Y(\infty)$,[3] it follows from Proposition 2.11 of Shaked and Shanthikumar (1988) that $Eh(Y(\infty))$ is a decreasing and convex function of μ.

We now fix $M > 0$ and establish, by induction, that if 1 and 2 hold at time $n = 0$ for $Y_1^M, \ldots, Y_4^M$, then they hold for all $n = 1, 2, \ldots$. For brevity, we omit the superscript M from the subsequent terms. In addition to 1 and 2 we define a third property as follows: $\tilde{1}$. $Y_1(n) + Y_4(n) = Y_2(n) + Y_3(n)$, that is, Property $\tilde{1}$ is Property 1 with an equality replacing the inequality. We first establish that if Properties $\tilde{1}$ and 2 are satisfied at time n, then they hold at time $n + 1$. Let $v = \lambda + \mu_4 + M$ be an upper bound on the total transition rate of the processes $Y_1, \ldots, Y_4$. For n, such that Properties $\tilde{1}$ and 2 hold, we define the following possible uniformized and coupled transitions:

Arrival + Balking. With probability λ/v we have a new order arriving into all four systems. When a new order arrives, it balks system i with probability $\beta(Y_i(n))$. This is done as follows: Let $Y_{(1)}(n) \ge Y_{(2)}(n) \ge Y_{(3)}(n) \ge Y_{(4)}(n)$ be the order statistics for $Y_i(n)$, $i = 1, \ldots, 4$. Respectively, refer to system (i) as the systems whose head count is $Y_{(i)}(n)$.

[3] This can be shown by writing down the stationary distributions of the corresponding birth and death processes explicitly and showing that those distributions converge to the limiting one, with unbounded reneging rates.

Note that from Properties $\tilde{1}$ and 2, it follows that $Y_{(1)}(n) = Y_1(n)$ and $Y_{(4)}(n) = Y_4(n)$. Now let $\beta_{(i)} = \beta(Y_{(i)}(n))$. From the monotonicity and concavity of $\beta(\cdot)$ it follows that: (a) $\beta_{(1)} \geq \beta_{(2)} \geq \beta_{(3)} \geq \beta_{(4)}$, and (b) $\beta_{(1)} + \beta_{(4)} \leq \beta_{(2)} + \beta_{(3)}$.

Now let $U \sim \text{Uniform}(0, 1)$. (i) If $U \leq \beta_{(4)}$, then balk in all four systems. (ii) If $U \leq \beta_{(2)} + \beta_{(3)} - 1$, then balk from queues (1), (2), and (3). (iii) If $U \leq \beta_{(3)}$, then balk in queues (3) and (1) only. (iv) If $U \leq \beta_{(1)}$, then balk in queues (2) and (1) only. (v) If $U \leq \beta_{(2)} + \beta_{(3)} - \beta_{(4)}$, balk in queue (2) only.

To verify that the balking occurs according to the right probabilities, note that in systems (1), (3), and (4) the balking probabilities are trivially equal to the required probabilities, provided that $\beta_{(4)} = \min\{\beta_{(i)}\}$ and $\beta_{(1)} \geq \beta_{(3)}$ both are guaranteed by a. In queue (2), if $\beta_{(2)} + \beta_{(3)} - \beta_{(4)} < 1$ balking will occur with probability: $\beta_{(4)} + (\beta_{(2)} + \beta_{(3)} - \beta_{(4)} - \beta_{(3)}) = \beta_{(2)}$, provided that $\beta_{(2)} + \beta_{(3)} - \beta_{(4)} \geq \beta_{(1)}$, which is equivalent to b. Similarly, if $\beta_{(2)} + \beta_{(3)} - \beta_{(4)} \geq 1$, balking in this queue will occur with probability: $(\beta_{(2)} + \beta_{(3)} - 1) + (1 - \beta_{(3)}) = \beta_{(2)}$.

Service Completion. With probability μ_4/v we have a service completion event. To determine which systems are going to indeed have service completions (as opposed to a transition from a state to itself), let $U \sim \text{Uniform}(0, 1)$. (a) If $U < \mu_1/\mu_4$, we have service completions from all systems for which $Y_i(n) > 0$. (b) If $\mu_1/\mu_4 \leq U < \mu_2/\mu_4$, we have departures in systems 2 and 4 only, whenever the corresponding queues are nonempty. (c) If $\mu_2/\mu_4 \leq U < 1$, we have departures in systems 3 and 4 only, whenever the corresponding queues are nonempty.

It is easy to see that system i has a service completion with probability μ_i/v as long as $Y_i(n) > 0$ (recall that $\mu_1 + \mu_4 = \mu_2 + \mu_3$). Note that the reason we do not simply have a service completion from system i whenever $U < \mu_i/\mu_4$ is that in this case we may have a service completion from system 4 only, which may violate Property 1.

A Reneging Job (Order Cancellation). Finally, with probability $[\eta(Y_i(n)) \wedge M]/v$ we have an order cancellation from system i. The coupling works as follows: Let $Y_{(1)}(n) \geq Y_{(2)}(n) \geq Y_{(3)}(n) \geq Y_{(4)}(n)$ be the ordered statistics of $Y_1(n), \ldots, Y_4(n)$, and let $\xi_{(i)} = \eta_M(Y_{(i)}(n)) = \min\{\eta(Y_{(i)}(n)), M\}$. Note that Property $\tilde{1}$ and the convexity of $\eta_M(\cdot)$ imply that $\xi_{(1)} + \xi_{(4)} \leq \xi_{(2)} + \xi_{(3)}$ (that is, the inequality with respect to the ξ_is is the opposite of Property 1). Let $U \sim \text{Uniform}(0, 1)$ be the random variable that determines the reneging from all systems. Let $m = \max\{M, \xi_{(3)} + \xi_{(2)} - \xi_{(4)}\}$. (a) If $U < \xi_{(4)}/m$, we will have one order cancellation from all the systems such that $Y_i(n) > 0$. (b) If $\xi_{(4)}/m \leq U < \xi_{(3)}/m$, we have one order cancellation from each of the systems (3) and (1) (provided that $Y_{(i)}(n) > 0$, for $i = 1, 3$). (c) If $\xi_{(3)}/m \leq U < \xi_{(1)}/m$, we have one order cancellation from each of the systems (2) and (1) (provided that $Y_{(i)}(n) > 0$, for $i = 1, 2$). (d) If $\xi_{(1)}/m \leq U < (\xi_{(3)} + \xi_{(2)} - \xi_{(4)})/m$, we have one order cancellation from system (2), provided that $Y_{(2)}(n) > 0$.

Note that, given this setup, an order cancellation occurs in system (i) with probability $[\eta(Y_{(i)}(n)) \wedge M]/v$.

This completes the proof of the theorem for the single server case. (Detailed verification of Properties 1 and 2 is given in the technical appendix.)

It is left to prove the theorem for the general multiserver ($S > 1$) case. To extend the above proof to the *M/M/S* system, the only case that needs to be treated is service completions. Let $Y_1 + Y_4 = Y_2 + Y_3$, $Y_1 \geq \max\{Y_2, Y_3, Y_4\}$, $\mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ and $\mu_1 + \mu_4 = \mu_2 + \mu_3$. Then we shall show that

$$\mu_1 \min\{Y_1, S\} + \mu_4 \min\{Y_4, S\} \leq \mu_2 \min\{Y_2, S\} + \mu_3 \min\{Y_3, S\}. \tag{4}$$

To see this consider the following cases: (1) If all Ys are greater than or equal to S, then (4) is true by assumption. (2) If only $Y_4 < S$, then, again, (4) is true by assumption. (3) If only Y_4 and Y_3 (wlog $Y_2 \geq Y_3$) are less than S, then if (4) is false, then

$$\mu_1 S + \mu_4 Y_4 > \mu_2 S + \mu_3 Y_3. \tag{5}$$

But

$$\mu_1 S + \mu_4 S \leq \mu_2 S + \mu_3 S, \tag{6}$$

and (5)–(6) yields $\mu_4(S - Y_4) < \mu_3(S - Y_3)$, which is false because $Y_4 \leq Y_3$ and $\mu_4 \geq \mu_3$. (4) If only Y_2, Y_3, Y_4 are less than S, then if (4) is false, then

$$\mu_1 S + \mu_4 Y_4 > \mu_2 Y_2 + \mu_3 Y_3. \tag{7}$$

However, (recall wlog $Y_2 \geq Y_3$) $\mu_1 Y_1 + \mu_4 Y_4 \leq \mu_1 Y_2 + \mu_4 Y_3$ (because we may increase Y_4 by an amount equal to the decrease in Y_1) $\leq \mu_2 Y_2 + \mu_3 Y_3$ (because we may increase μ_1 by same amount as the decrease in μ_4). Thus,

$$\mu_1 Y_1 + \mu_4 Y_4 \leq \mu_2 Y_2 + \mu_3 Y_3. \tag{8}$$

Then (7) and (8) lead to a contradiction. (5) If all Ys are less than S, then (4) follows from (8).

Finally, (4) implies that we can couple the four systems such that the second and third have more service completions on each sample path and that Properties $\tilde{1}$ and 2 hold at each service completion.

Now consider the number of servers. Fix μ and suppose that $0 < S_1 \leq S_2 \leq S_3 \leq S_4$, and $S_1 + S_4 = S_2 + S_3$. We show that there exist versions of the corresponding head-count processes such that for all $t \geq 0$:

1. $Y_1(t) + Y_4(t) \geq Y_2(t) + Y_3(t)$. a.s., and
2. $Y_1(t) \geq \max\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s.

The proof is identical to that for μ as the decision variable, except for the service completion step. Here, too, we have to verify that if Properties 1 or 2 holds with equality at time n, both of them are still true at time $n+1$. Let $Z_i = S_i \wedge Y_i$; then Z_i is the number of busy servers in system i, $i = 1, 2, 3, 4$. We claim that if $Y_1(n) + Y_4(n) = Y_2(n) + Y_3(n)$, then

$$Z_1(n) + Z_4(n) \leq Z_2(n) + Z_3(n), \tag{9}$$

and that if $Y_1(n) = Y_i(n)$ for some $i = 2, 3$, or 4, then

$$Z_1(n) \le Z_i(n). \tag{10}$$

Both (9) and (10) can be easily verified by examining all the different possible combinations of $Z_i = S_i$ or $Z_i = Y_i$, for $i = 1, 2, 3, 4$. Now that we have (9) and (10) we can couple service completions in a way that Properties 1 and 2 will not be violated: Suppose that Property 1 holds with an equality. We claim that either $Z_1 = \min\{Z_i\}$ or $Z_4 = \min\{Z_i\}$, $i = 1, 2, 3, 4$. To see that, suppose that $Z_2 = \min\{Z_i\}$. This implies that $Y_2 \le S_2$, because $S_2 \ge S_1 \ge Z_1$. But in this case, $Y_4 \le Y_2 \le S_2 \le S_4$. In particular, $Z_4 = Y_4 = \min\{Z_i\}$. Similarly, we can argue that if $Z_3 = \min\{Z_i\}$, then $Z_4 = Z_3 = \min\{Z_i\}$. Suppose that $Z_1 = \min\{Z_i\}$, then:

1. With probability $Z_1/(Z_2 + Z_3)$ we have departures out of all systems.

2. With probability $(\min\{Z_2, Z_4\} - Z_1)/(Z_2 + Z_3)$ we have departures from systems 4 and 2 only.

3. With probability $(Z_4 - Z_2)^+/(Z_2 + Z_3)$ we have departures from systems 4 and 3 only.

4. With probability $(Z_2 - Z_4)^+/(Z_2 + Z_3)$ we have a departure from system 2 only.

5. Finally, with probability $(Z_3 - Z_1 - (Z_4 - Z_2)^+)/(Z_2 + Z_3)$ we have a departure from system 3 only.

6. Notice that with probability $Z_1/(Z_2 + Z_3)$ no transition occurs. If $Z_4 = \min\{Z_i\}$, then the transitions parallel the ones above, switching the roles of systems 1 and 4.

Two important observations with respect to the above transitions are that (a) all departures from systems 1 or 4 are coupled with departures from systems 2 or 3, and (b) if $Y_i = Y_1$ for some $i = 2, 3$, or 4, then one can verify that in this case, because of (10), any departure from system 1 will be coupled with a departure from system i. Therefore, the induction step is confirmed under service completion transitions, which completes the proof of the theorem. □

Proof of Proposition 1. The proof is a straightforward application of Theorem 1. Details are given in the technical appendix. □

Proof of Theorem 2. The proof of Theorem 2 for μ as the decision variable is based on three propositions. The first (Proposition 1) shows that $C(\mu, S; \lambda, \eta)$ is convex in μ; therefore, for all given values of λ and η, $\mu^*(\lambda, \eta)$ is well defined (although it may be non-unique), and any local minimum of $C(\mu, S; \lambda, \eta)$ is also a global minimum. (Proposition 1 establishes convexity of the cost in S as well.) The second proposition (Proposition 3) shows that $\mu^*(\lambda, \eta)$ is increasing in λ. (A similar result can be established for S as a decision variable—see Proposition 4.) Finally, in the third proposition (Proposition 5) we establish that $\mu^*(\lambda, \eta)$ may either increase or decrease in η. (That the optimal S may either increase or decrease in S is shown through the numerical example presented in Figure 3.) □

Building on the concept of *sample-path convexity*, we define the *sample-path submodularity* property that implies monotonicity of the expected-cost-minimizing value of one parameter in a second parameter.

Definition. Let $X = X_{\gamma,\delta}$ be a stochastic process that depends on the two parameters γ and δ. We say that X is *pathwise submodular* with respect to γ and δ if for all $\gamma_L < \gamma_H$ and $\delta_L < \delta_H$ we have four processes $\hat{X}_{\gamma,\delta}$, $\gamma = \gamma_L, \gamma_H$, $\delta = \delta_L, \delta_H$ that are defined on the same probability space, such that:

1. $\hat{X}_{\gamma,\delta}$ is a version of $X_{\gamma,\delta}$ for every fixed pair (γ, δ) (that is, $\hat{X}_{\gamma,\delta} \stackrel{st}{=} X_{\gamma,\delta}$), and

2. $\hat{X}_{\gamma_H,\delta_H} - \hat{X}_{\gamma_H,\delta_L} \le \hat{X}_{\gamma_L,\delta_H} - \hat{X}_{\gamma_L,\delta_L}$, a.s.

The next theorem establishes the connection between the sample-path submodularity property and monotonicity.

Theorem 4. *Let $X = X_{\gamma,\delta}$ be a stochastic process, and let $g(\gamma, \delta) = EX_{\gamma,\delta}$ be its expected value in steady state. Suppose that $g(\cdot)$ is convex in γ for every fixed δ and that X is pathwise submodular with respect to these two variables. Let $\gamma^*(\delta)$ be the (possibly non-unique) value of γ that minimizes $g(\gamma, \delta)$ for every fixed value of δ; then $\gamma^*(\delta)$ is increasing in δ.*

Proof. First note that it is straightforward to show that if X is pathwise submodular in δ and γ, and the steady state of X exists, then $g(\cdot)$ is submodular in these two variables. Let δ_L, δ_H be two values of δ such that $\delta_L < \delta_H$. Let $\gamma^*(\delta_L)$ be a value of γ that minimizes $g(\gamma, \delta_L)$. We need to show that there is $\hat{\gamma} \ge \gamma^*(\delta_L)$ such that $\hat{\gamma}$ minimizes $g(\gamma, \delta_H)$. By contradiction, assume that for all optimal solutions $\gamma^*(\delta_H)$ of $g(\gamma, \delta_H)$, we have $\gamma^*(\delta_H) < \gamma^*(\delta_L)$. In particular,

$$\begin{aligned} 0 &\le g(\gamma^*(\delta_H), \delta_L) - g(\gamma^*(\delta_L), \delta_L) \\ &\le g(\gamma^*(\delta_H), \delta_H) - g(\gamma^*(\delta_L), \delta_H) \le 0, \end{aligned} \tag{11}$$

where the first inequality follows from the optimality of $\gamma^*(\delta_L)$, the second one follows from the sample-path submodularity and the assumption that $\gamma^*(\delta_H) < \gamma^*(\delta_L)$, and the third follows from the optimality of $\gamma^*(\delta_H)$. In particular, (11) implies that $g(\gamma^*(\delta_H), \delta_H) = g(\gamma^*(\delta_L), \delta_H)$, which in turn implies that $\gamma^*(\delta_L)$ minimizes $g(\gamma, \delta_H)$. This leads to a contradiction. □

The next proposition establishes that the head-count process is pathwise submodular in μ and λ (η will be omitted from the current expressions for expository purposes). From Theorem 4 it then follows that $\mu^*(\lambda)$ is nondecreasing in λ.

Proposition 3. *Suppose that the reneging rate $\eta(y) = \eta y$ for some fixed $\eta \ge 0$. For any values of λ and μ, let $Y_{\lambda,\mu}$ represent the head-count process when the arrival rate is λ and the service capacity is μ. Then $Y_{\lambda,\mu}$ is pathwise submodular in λ and μ.*

Note that the proposition only establishes the pathwise submodularity of $Y_{\lambda,\mu}$. However, it is straightforward to verify that this implies the sample-path submodularity of $\eta Y_{\lambda,\mu} + k\mu S$ in these two parameters. The proof of Proposition 3 follows the sample path approach. More specifically, we show that for all $\lambda_L < \lambda_H$ and $\mu_L < \mu_H$ there exist versions of $Y_{\lambda,\mu}$ for $\lambda \in \{\lambda_L, \lambda_H\}$ and $\mu \in \{\mu_L, \mu_H\}$ such

that the following three properties hold at all times $t \geq 0$: (i) $Y_{\lambda_H,\mu_L}(t) = \max[Y_{\lambda,\mu}(t)\colon \lambda \in \{\lambda_L, \lambda_H\}, \mu \in \{\mu_L, \mu_H\}]$, a.s., (ii) $Y_{\lambda_L,\mu_H}(t) = \min[Y_{\lambda,\mu}(t)\colon \lambda \in \{\lambda_L, \lambda_H\}, \mu \in \{\mu_L, \mu_H\}]$, a.s., and (iii) $Y_{\lambda_H,\mu_H}(t) - Y_{\lambda_H,\mu_L}(t) \leq Y_{\lambda_L,\mu_H}(t) - Y_{\lambda_L,\mu_L}(t)$, a.s.

Similarly to the proof of Theorem 1, we show (i), (ii), and (iii) using time discretization and uniformization, and then establishing these properties using sample-path coupling and induction on time. Details are given in the technical appendix.

Proposition 4. *Suppose that the reneging rate $\eta(y) = \eta y$ for some fixed $\eta \geq 0$. For any values of λ and S, let $Y_{\lambda,S}$ represent the head-count process when the arrival rate is λ and the number of servers is S. Then $Y_{\lambda,S}$ is pathwise submodular in λ and S.*

The proof is analogous to the proof of Proposition 3. Details are given in the technical appendix.

Proposition 5. *Fix the values of S and λ, suppose that the reneging rate $\eta(y) = \eta y$ and let $\mu^*(\eta)$ be the optimal service rate that minimizes the cost function $C(\mu, S; \lambda, \eta)$. Then $\mu^*(\eta)$ can either increase or decrease in η.*

Proof. Recall the cost function $C(\mu; \eta) := C(\mu, S; \lambda, \eta) = \eta EY(\infty) + k\mu S$. Suppose that $S = 1$. To prove the proposition we first show that for arbitrary values of λ and k with $0 < k < 1$, $\mu^*(\eta)$ may decrease in η. To show that, we note that the definitions $C(\mu; \eta = 0) = (\lambda - \mu)1_{\{\mu \leq \lambda\}} + k\mu$, and $C(\mu; \eta = \infty) = \lambda + k\mu$ are continuous extensions of the cost function $C(\cdot)$ for all values η in the closed interval $[0, \infty]$. However, notice that $\mu^*(\eta = 0) = \lambda$, whereas, $\mu^*(\eta = \infty) = 0$; that is, $\mu^*(\eta)$ may decrease with η.[4]

To show that $\mu^*(\eta)$ may also increase in η, all we have to show is that there exist $0 < k < 1$ and $\eta_k > 0$ such that $\mu_k^*(\eta_k) > \lambda$ (recall that $\mu^*(\eta = 0) = \lambda$). We show, in fact, that a stronger result applies—namely, that for every fixed value of $\eta > 0$ there exists a value $k = k(\eta)$, $0 < k < 1$ such that $\mu_k^*(\eta) > \lambda$, where $\mu_k^*(\eta)$ stands for the optimal capacity that minimizes the cost function $C(\mu; \lambda, \eta) = \eta EY(\infty) + k\mu$. To show that, fix the value of $\eta > 0$ and note that the function $f(\mu) = \eta EY(\infty)$ is decreasing and convex in μ (Theorem 1). We claim that it is sufficient to show that:

There exists μ_0 such that:

$$\mu_0 > \lambda,\ f(\mu_0) < f(\lambda) \text{ and } f'(\mu_0^-) > -1, \tag{12}$$

where $f'(\mu_0^-)$ is the directional derivative of f at $\mu = \mu_0$ from below (exists because of Lemma 3.1.5 of Bazaraa et al. 1993). If (12) is true, then the convexity of $f(\mu)$ implies that $f'(\mu_0^-) \leq f'(\mu_0^+)$ (here, $f'(\mu_0^+)$ is the directional derivative of f at $\mu = \mu_0$ from above). Let k be such that $f'(\mu_0^-) \leq -k \leq f'(\mu_0^+)$; then $C'(\mu_0^-) = f'(\mu_0^-) + k \leq 0$, and $C'(\mu_0^+) = f'(\mu_0^+) + k \geq 0$. In particular, $\mu_k^*(\eta) = \mu_0$ is a local minimum for $C(\cdot)$, and from convexity, it is also a global minimum.

To establish (12), note that from flow conservation $f(\mu) = \eta EY(\infty) = \lambda - \mu P(Y(\infty) > 0)$. In particular, $f(\mu = \lambda) > 0$, and $\lim_{\mu\to\infty} f(\mu) = 0$. Because $f(\mu)$ is a nonincreasing function of μ, this implies that there exists $\mu_1 > \lambda$ such that $f(\mu) < f(\lambda)$, for all $\mu \geq \mu_1$. Now note that if $f'(\mu^-) \leq -1$ for all $\mu \geq \mu_1$, then $f(\mu) < 0$ for μ large enough, which is a contradiction. This shows that μ_0 is well defined. □

Proof of Theorem 3. First consider the parameter region $\mu \geq \eta$ and the case of μ as decision variable. Following the notation of the proof of Theorem 1, let $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ be four service rates such that $\mu_1 + \mu_4 = \mu_2 + \mu_3$. Assume that the reneging rate η is bounded above by μ_4. This is a weaker condition than the condition $\mu \leq \eta$ stated in the theorem, but it turns out to be sufficient to establish its results.

Analogously to the proof of Theorem 1, let $Y_1, \ldots, Y_4$ be discretized and uniformized versions of the head count with service rates $\mu_1, \ldots, \mu_4$, respectively, that satisfy properties:

1. $Y_1(n) + Y_4(n) \geq Y_2(n) + Y_3(n)$, a.s. and
2. $Y_1(n) \geq \max\{Y_2(n), Y_3(n), Y_4(n)\}$, a.s. at time $n = 0$.

By induction, we wish to show that Properties 1 and 2 hold for all $n \geq 0$.

The induction proof of Properties 1 and 2 goes through by the simple construction explained next. Note that arrivals and service completions do not introduce a problem. For reneging, we use a slightly different method involving a transfer of a customer. The transfer need not occur physically, but it helps to visualize why we need the condition that $\eta \leq \mu_4$. (We could use the same proof used earlier without this device.) Specifically, transferring a customer from systems 1 to 4 is equivalent to comparing the current set of systems with another set that has $Y_1 - 1$ and $Y_4 + 1$ customers, which in turn is comparable to systems 2 and 3. The idea is that because Y_4 is smaller than Y_1, if we transfer a customer from system 1 to system 4, it will change the total departure rate in both systems combined in the desired direction (either through service completions or through reneging). For example, if both systems have idle servers, then the combined rate of departures due to service will increase. If only system 4 has an idle server, then the rate will increase because the reneging rate is smaller than μ_4. Finally, if both are busy, then the total reneging rate remains the same.

The details of the customer transfer are as follows: One can transfer customers from systems 1 to 4 until one of two events happens—either Y_4 equals the minimum of Y_2 and Y_3, or Y_4 equals S. In the first case, after the transfer, Y_1 will equal the maximum of Y_2 and Y_3. In the second case all systems will have S or more customers. The transfer will not decrease the rate at which queues deplete in systems 1 and 4 because of the assumption on the reneging rate. Moreover, Properties $\tilde{1}$ (or 1) and 2 will continue to

[4] The continuity of $\mu^*(\eta)$ (which follows from Theorem 3.1.3 of Bazaraa et al. 1993 and the implicit functions theorem) may be used to show that $\mu^*(\eta)$ indeed decreases for some points on the interval $(0, \infty)$.

hold. It thus follows that the induction proof goes through after this modification. In detail, in the first case the two sets of systems will have equal reneging rates. In the second case, $(Y_1 - S) + (Y_4 - S) = (Y_2 - S) + (Y_3 - S)$. The reneging rates depend on these four quantities and the earlier proof for Theorem 1 goes through.

With respect to μ as a decision variable, it is left to establish that Y is stochastically decreasing in μ for $\mu \leq \eta$. This proof is a trivial application of the sample-path coupling approach. Details are omitted. The detailed verification of the induction step for S as a decision variable is given in the technical appendix. □

Proof of Proposition 2. The proof is a straightforward application of Theorem 3. Details are given in technical appendix. □

References

Afeche, P. 2004. Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delaying tactics. Working paper, University of Toronto, Canada.

Armony, M., M. Haviv. 2003. Price and delay competition between two service providers. *Eur. J. Oper. Res.* **147**(1) 32–50.

Armony, M., E. L. Plambeck. 2005. The impact of duplicate orders on demand estimation and capacity investment. *Management Sci.* **51**(10) 1505–1518.

Baccelli, F., G. Hebuterne. 1981. On queues with impatient customers. F. J. Kylatra, ed. *Performance '81*. North-Holland Publishing Company, Amsterdam, 159–179.

Bazaraa, M. S., H. D. Sherali, C. M. Shetty. 1993. *Nonlinear Programming: Theory and Algorithms*, 2nd ed. John Wiley & Sons, Inc., New York.

Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Brandt, A., M. Brandt. 2002. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/S + GI$ system. *Queueing Systems* **41** 73–94.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100**(469) 36–50.

Chen, L., E. L. Plambeck. 2008. Dynamic inventory management with learning about the demand distribution and substitution probability. *Manufacturing Service Oper. Management* **10**(2) 236–256.

Cohen, M., P. Kleindorfer, H. Lee. 1988. Service constrained (s, S) inventory systems with priority demand classes and lost sales. *Management Sci.* **34**(4) 482–499.

Duenyas, I., W. J. Hopp. 1995. Quoting customer lead times. *Management Sci.* **41** 43–57.

Fairlie, R. 2004. How a custom PC can come with some very alien payment policies. *Computer Shopper* (March 10) 28.

Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54**(2) 324–338.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4**(3) 208–227.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Oper. Management* **7**(1) 20–36.

Ho, T.-H., S. Savin, C. Terwiesch. 2002. Managing demand and sales dynamics in new product diffusion. *Management Sci.* **48**(2) 187–206.

Hopp, W. J., M. R. Sturgis. 2001. A simple and robust leadtime-quoting policy. *Manufacturing Service Oper. Management* **3**(4) 331–336.

Johansen, S. G., R. M. Hill. 2000. The (r, Q) control of a periodic-review inventory system with continuous demand and lost sales. *Internat. J. Production Econom.* **68** 279–286.

Johansen, S. G., A. Thorstenson. 1993. Optimal and approximate (Q, r) inventory policies with lost sales and gamma-distributed leadtimes. *Internat. J. Production Econom.* **30–31** 179–194.

Johansen, S. G., A. Thorstenson. 1996. Optimal (r, Q) inventory policies with Poisson demands and lost sales: Discounted and undiscounted cases. *Internat. J. Production Econom.* **46–47** 359–371.

Kapuscinski, R., S. Tayur. 2007. Reliable due date setting in a capacitated MTO system with two customer classes. *Oper. Res.* **55**(1) 56–74.

Keskinocak, P., R. Ravi, S. Tayur. 2001. Scheduling and reliable lead time quotation for orders with availability intervals and lead time sensitive revenues. *Management Sci.* **47**(2) 264–279.

Koole, G., A. Pot. 2006. A note on profit maximization and monotonicity for inbound call centers. Working paper, VU University of Amsterdam, The Netherlands.

Kumar, S., J. Swaminathan. 2003. Diffusion of innovations under supply constraints. *Oper. Res.* **51**(6) 866–879.

Lederer, P. J., L. Li. 1997. Pricing, production, scheduling, and delivery-time competition. *Oper. Res.* **45**(3) 407–420.

Li, L., Y. S. Lee. 1994. Pricing and delivery-time performance in a competitive environment. *Management Sci.* **40**(5) 633–646.

Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Sci.* **49**(8) 1018–1038.

Mandelbaum, A., N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36**(1–3) 1084–1134.

Mandelbaum, A., S. Zeltyn. 1998. Estimating characteristics of queueing networks using transactional data. *Queueing Systems* **29** 75–127.

Mandelbaum, A., S. Zeltyn. 2006. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. Working paper, Technion, Israel Institute of Technology, Haifa, Israel.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* **38**(5) 870–883.

Moinzadeh, K., S. Nahmias. 1988. A continuous review model for an inventory system with two supply modes. *Management Sci.* **6** 761–773.

Nahmias, S. 1979. Simple approximations for a variety of dynamic leadtime lost-sales inventory models. *Oper. Res.* **27**(5) 904–924.

Plambeck, E. L. 2004. Optimal leadtime differentiation via diffusion approximations. *Oper. Res.* **52**(2) 213–228.

Plambeck, E. L., A. R. Ward. 2006. Optimal control of a high-volume assemble-to-order system. *Math. Oper. Res.* **31**(3) 453–477.

Plambeck, E. L., A. R. Ward. 2007. Note: A separation principle for a class of assemble to order systems with expediting. *Oper. Res.* **55**(3) 603–609.

Porteus, E. L. 2002. *Foundations of Stochastic Inventory Theory*. Stanford University Press, Palo Alto, CA.

Randhawa, R., S. Kumar. 2005. Multi-server loss systems with subscribers. Working paper, Stanford University, Palo Alto, CA.

Reed, J. E., A. R. Ward. 2008. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Math. Oper. Res.* **33**(3) 606–644.

Shaked, M., J. G. Shanthikumar. 1988. Stochastic convexity and its applications. *Adv. Appl. Probab.* **20** 427–446.

Shanthikumar, J. G., D. D. Yao. 1989. Second-order stochastic properties in queueing systems. *Proc. IEEE* **77** 162–170.

Van Mieghem, J. 2000. Price and service discrimination in queueing systems: Incentive compatibility of Gcμ scheduling. *Management Sci.* **46**(9) 1249–1267.

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized cμ rule. *Ann. Appl. Probab.* **5**(3) 809–833.

Ward, A. R., P. Glynn. 2005. A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems* **50** 371–400.

Ward, K. 2006. Asymptotically optimal admission control of a queue with impatient customers. Working paper, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

Whitt, W. 2005. Engineering solution of a basic call-center model. *Management Sci.* **51**(2) 221–235.

Whitt, W. 2006a. Fluid models for multiserver queues with abandonments. *Oper. Res.* **24**(1) 37–54.

Whitt, W. 2006b. Staffing a call center with uncertain arrival rate and absenteeism. *Production Oper. Management* **15**(1) 88–102.

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