MANAGEMENT SCIENCE
Vol. 52, No. 11, November 2006, pp. 1778–1791
ISSN 0025-1909 | EISSN 1526-5501 | 06 | 5211 | 1778

DOI 10.1287/mnsc.1060.0587


# Dynamic Control of an M/M/1 Service System with Adjustable Arrival and Service Rates

Barış Ata
Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208, b-ata@kellogg.northwestern.edu

Shiri Shneorson
Intel Corporation, 2220 Mission College, Santa Clara, California 95054, shiri@stanfordalumni.org

We study a service facility in which the system manager dynamically controls the arrival and service rates to maximize the long-run average value generated. We initially consider a rate-setting problem where the service facility is modeled as an M/M/1 queue with adjustable arrival and service rates and solve this problem explicitly. Next, we use this solution to study a price-setting problem, where customers are utility maximizing and price- and delay-sensitive, and the system manager chooses state-dependent service rates and prices. We find that the optimal arrival rate is decreasing and the optimal service rate is increasing in the number of customers in the system; however, the optimal price need not be monotone. We also show that under the optimal policy, the service facility operates as one with a finite buffer. Finally, we study a numerical example to compare the social welfare achieved using a dynamic policy to that achieved using static policies and show the dynamic policy offers significant welfare gains.

*Key words*: stochastic models of service systems; dynamic control; delay-sensitive customers
*History*: Accepted by Michael Fu, stochastic models and simulation; received July 30, 2004. This paper was with the authors $3\frac{1}{2}$ months for 4 revisions.

## 1. Introduction

In the highly competitive business environment prevalent today, the ability to dynamically control cost and value factors can give an operation a significant competitive advantage. The control of value through dynamic pricing is now a common practice in industries such as the airline industry. Merits of such dynamic control mechanisms are often the focus of the revenue management literature (cf. Talluri and Van Ryzin 2004). It is less common to control the cost factors dynamically, at least partly because it is often hard to implement. Still, the potential advantages it could offer are significant; therefore, many businesses have experimented with dynamic control of cost factors in recent years. *Computing on demand* illustrates such a strategy. Computing on demand means that a company purchases either server time, storage space, or processing power (or a combination) from another company according to experienced or anticipated demand, instead of investing in a fixed capacity, in-house computing infrastructure. This enables companies to rapidly adjust their computing capacity to spikes in usage while keeping their IT spending relatively low. The companies that provide the service, in return, charge their customers the same way an electric company does: When usage spikes, so does the bill.[1]

*Bandwidth on demand* is another example of a strategy that enables dynamic cost control. There, rather than contracting for a fixed bandwidth Internet connection, companies have a flexible contract with their Internet service provider that allows them to change their bandwidth requirement and pay based on actual usage. Typically, companies pay a fixed fee to reserve with their Internet service providers the right to increase or decrease the bandwidth of their Internet link and then pay a variable fee according to demand.[2]

Although dynamic capacity (cost) control is relatively easy to implement in IT services, where changing the amount of computing power assigned to a customer (computing on demand) or a customer's bandwidth quota (bandwidth on demand) has both low setup and marginal costs, the trend toward flexible operations is also evident in more

[1] IBM's computing-on-demand services today serve big companies such as Exxon Mobil's Mobil Travel Guide and American Express.

[2] Bandwidth on demand service is offered by leading Internet service providers such as AT&T.

traditional industries. For example, a call center that serves multiple types of customers, each calling a different dedicated phone line, often trains its agents to handle more than one customer type to enable a dynamic allocation of agents to calls based on real-time demand.[3]

In all settings where dynamic cost and value control are feasible, the tradeoff between the potential gain on the one hand and the implementation complexity and cost on the other hand plays a key role when making the adoption decision. Managing a service facility is a complex task on its own; therefore, having the facility serve multiple customers at varying prices and allocating the capacity dynamically on demand is a much more challenging task. In our work, we address this challenge. We analyze a stylized model of a service facility that operates as an M/M/1 queue and find the optimal state-dependent capacity and price decisions for a system manager who maximizes the total system welfare. Then, we compare the gains the optimal dynamic policy offers over optimal static policies.

A canonical setting where our model is arguably most appropriate is an organization such as a firm consisting of an internal service department and several user departments. The provision of services to competing user departments through a shared service facility is often characterized by the presence of negative externalities: When a user expands his usage of the limited capacity, he imposes costs on the rest of the system in the form of degraded service quality. Following the classical approach of internal pricing (cf. Hirshleifer 1964), our solution facilitates the socially optimal dynamic arrival and service rates, which maximize the system welfare, through dynamic pricing.

Undoubtedly, in most real-world cases, service facilities' queue dynamics are more sophisticated than that of the M/M/1 queue. However, we use this stylized Markovian model in our analysis as an approximation to real-world systems because of its tractability. Arguably, the M/G/1 queue would be a better model, but it is not amenable to dynamic analysis. Although the M/M/1 model may not capture all aspects of the system behavior, it captures many of the congestion-related phenomenon. Moreover, using this model, we are able to explicitly characterize the optimal policy, learn about its key properties, and evaluate the added value of deploying a dynamic policy.

We assume that service requests arrive at the facility according to a Poisson process, and that each such request has an exponentially distributed service time requirement. The system manager can control dynamically both cost and value factors, and the control decision is made every time the state of the system changes, either because of a new arrival or a termination of service. The system manager's objective is to maximize the system's *long-run average welfare*, which is defined as the value from service minus capacity costs and holding costs, where the latter are simply customers' delay costs. We assume that the customers requesting the service are utility-maximizing and delay-sensitive decision makers. Each such customer enjoys a random utility from receiving the service but, at the same time, incurs a cost that is proportional to the length of delay experienced while waiting for the service to be completed. Thus, a customer will decide to submit a service request if her expected net utility (that is, her utility from the service minus delay cost and service fee) is positive.

The cost factor being controlled by the system manager is the service rate at which the system operates. The *service rate* is defined as the number of average service requests the system can process per unit of time. For the value factor, our goal is to find the set of prices that the system manager should post to maximize the welfare generated by the system. As a first step to finding these optimal prices, we study a rate-setting problem in which we disregard the customers' decisions and prices and assume that the system manager can directly control the arrival rates to the system and treat delay costs incurred by the customers as a holding cost that the system manager incurs. Then, we use the solution of the rate-setting problem to find optimal state-dependent prices a system manager should set to coordinate the arrival rates to those that maximize the welfare. That is, the system manager dynamically sets the prices that customers pay to use the service. Given these prices and information about expected system delay, customers decide whether or not to submit their service requests.

Focusing initially on the rate-setting problem, we find an explicit solution for the system manager's optimal choices of service and arrival-rate control. We show that the system manager's optimal choices of service and arrival rates are monotone in the state of the system. Specifically, the service rate is increasing and the arrival rate is decreasing with the number of requests waiting for service. It also turns out that, under this optimal policy, the service facility is equivalent to a system with a finite buffer, where the buffer size is characterized explicitly by our solution. That is, our solution of the rate-setting problem specifies the minimal "waiting-room" size (buffer size) required to implement the optimal strategy. We next consider the price-setting problem where the system manager is restricted to price setting (rather than arrival-rate setting). We find the set of prices and service rates that coordinate the system to same system optimality that was found for the rate-setting

[3] As an example, the reader may think of health care providers that offer dedicated phone lines to their major institutional clients.

problem, assuming that customers are well informed about their expected delays in the system.

We show, using a numerical example, that the optimal prices need not be monotone with the state of the system. That is, it might be optimal to charge low prices to customers arriving at a highly congested system, despite the need to decrease the arrival rate. Finally, we perform a numerical study to evaluate the social welfare gained by using dynamic policies over static ones. We observe that the welfare gain may be significant, and that it increases with the customers' delay costs. We also conclude that dynamic policies are most attractive in systems with small buffer sizes and, thus, with fewer possible states.

To repeat, our primary goal is to solve the price-setting problem, and the rate-setting problem is solved to facilitate that solution, although it could be of interest on its own right. In what follows, in §2, we briefly review the related literature. In §3, we present and solve the rate-setting problem. In §4, we solve the price-setting problem. Section 5 presents our numerical analysis. Section 6 presents comparative statics. Section 7 contains a summary and concluding remarks. Auxiliary proofs and derivations of comparative statics are provided in an online technical appendix available on the *Management Science* website at http://mansci.pubs.informs.org/ecompanion.html.

## 2. Literature Review

Our work combines, and is inspired by, two disparate streams of research. The first framework, often referred to as the *congestion pricing* literature, advocated by Naor (1969), Mendelson (1985), Dewan and Mendelson (1990), Mendelson and Whang (1990), Westland (1992), and others, provides a model of a service system where the customers requesting service are rational, utility-maximizing, and delay-sensitive decision makers. For example, Mendelson (1985) and Dewan and Mendelson (1990) study an internal service facility in an organization where a system manager sets both prices and capacities to maximize welfare under various delay cost structures. The controls analyzed in these two papers are static. That is, the service rate and the pricing decisions are made once, based on average performance estimates before any realization is observed, and does not change thereafter. Naor (1969) models a fixed capacity service system in which the arriving customers observe the queue length before deciding whether or not to enter the queue. However, the proposed system tolls, set to maximize either welfare or revenue, are static. That is, they do not depend on the state of the system. Mendelson (1985) also provides a framework to model behavior of customers who are sensitive to both price and delay.

The second stream of research that inspires our work is *dynamic control of queuing systems*, which has been studied extensively in the operations research literature in the past 30 years (see Stidham 1988, 2002 for surveys). In particular, two closely related papers to our work are George and Harrison (2001) and Low (1974). George and Harrison study a Markovian queue in which the system manager chooses the service rate dynamically to minimize the long-run average cost but has no control over the customers' arrival rate. Low (1974), on the other hand, studies the optimal control of a Markovian queue with a finite buffer in which the system manager controls the arrival rates through prices but has no control over the capacity level. The arriving customers, or jobs, are independent, rational decision makers. They observe the prices set by the system manager and decide whether or not to join the queue based on their net utilities. The system manager's goal in setting the prices is to maximize long-run average reward earned by serving customers. It turns out that the optimal prices advertised by the system manager are a nondecreasing function of the number of customers in the system.

Loosely speaking, our work combines the models of George and Harrison (2001) and Low (1974) and provides a new method to solve the combined problem (that is, the rate-setting problem) explicitly. More important, we generalize the dual control problem of value and cost factors of Mendelson (1985) and Dewan and Mendelson (1990) to a dynamic setting. In addition to explicitly characterizing the optimal controls in both the rate-setting and price-setting problems, we show that optimal prices need not be monotone in the state of the system and that the welfare gained by using dynamic policies over static ones can be significant.

Masuda and Whang (1999) study the dynamic pricing control problem for a fixed-capacity communication network in which the system manager wishes to maximize welfare and customers are delay-sensitive. The authors assume that the system manager has limited information about the demand curves and uses observed system performance as input for the pricing formula. They find that the maximum social welfare solution is an equilibrium and show that an instability problem arises because the system may never reach the desired equilibrium.

Paschalidis and Tsitsiklis (2000) and Paschalidis and Liu (2002) study dynamic pricing of a communication network that serves multiple customer classes. In this model, the prices set by the service provider may depend on the level of congestion, but capacity is fixed and the network operates as a loss network. In this context, the authors show that the performance of an optimal dynamic strategy is closely matched by a suitably chosen class-dependent static price.

Yoon and Lewis (2004) explore the problem of dynamic pricing and admission control in another interesting direction. The authors study a problem where the arrival and service rates are nonstationary and customers are price-sensitive. They establish several structural properties of the optimal policy, including the monotonicity of the optimal prices in the state of the system under various cost structures. The authors also propose a practical pointwise stationary approximation to the optimal dynamic policy.

Finally, another closely related paper is by Maglaras and Zeevi (2003), who address the problem of joint static pricing and capacity sizing using large capacity asymptotics for systems with shared resources in heavy traffic.

## 3. The Rate-Setting Problem

In this section, we study the problem of choosing the arrival and service rates as functions of the state of the system to maximize long-run average social welfare generated per unit of time by the service facility. The system manager chooses $\mu_n \in A$ and $\lambda_n \in B$ for $n = 0, 1, 2, \ldots,$ where $A = [0, M]$ and $B = [0, \Lambda]$, and the state of the system (that is, the queue length) evolves as a birth-and-death process with state-dependent arrival rates $\lambda_n$ and state-dependent service rates $\mu_n$ for $n = 0, 1, 2, \ldots$. Also given are two functions $c(\cdot)$ on $A$ and $b(\cdot)$ on $B$, where $c(\mu)$ is the cost rate associated with service rate $\mu$, and $b(\lambda)$ is the value rate associated with the arrival rate $\lambda$. It is assumed that $c(\cdot)$ is nondecreasing and convex on $A$ with $c(0) = 0$, and $b(\cdot)$ is increasing, strictly concave, continuously differentiable[4] on $B$ with $b(0) = 0$. In addition, the system manager incurs a holding cost of $h_n$ per unit of time when the state of the system is $n$. This corresponds to delay costs incurred by customers in the price-setting problem.[5] We assume that $h_n$ is nondecreasing in queue length with $h_0 = 0$ and $\lim_{n\to\infty} h_n = \infty$, which is a natural assumption in most settings including the price-setting problem we study in this paper. Finally, it is crucial for our analysis to assume that $b'(0) < \infty$, which derives the conclusion that under the optimal policy the service facility operates as one with a finite buffer. In the context of the price-setting problem, this means that customer valuations are bounded. It is natural to assume that customer valuations take values in a bounded interval, say $[\underline{u}, \bar{u}]$, in which case it follows easily that $b'(0) = \bar{u} < \infty$ (cf. §4).

[4] Assuming $b(\cdot)$ is differentiable gives rise to an appealing interpretation in the context of the price-setting problem. However, assumptions on both $b(\cdot)$ and $c(\cdot)$ can be further relaxed along the lines of George and Harrison (2001).

[5] In that context, holding cost is linear in the number of customers in the system. However, the solution to the rate-setting problem admits more general holding cost structures.

For us, a policy is a pair of vectors $(\vec{\lambda}, \vec{\mu})$, where $\vec{\lambda} = (\lambda_0, \lambda_1, \ldots)$ with all components belonging to the set $B$ and $\vec{\mu} = (\mu_1, \mu_2, \ldots)$ with all components belonging to the set $A$, assuming by convention that $\mu_0 = 0$. To be specific, we assume that the system is empty initially. The system manager's problem is to choose a policy $(\vec{\lambda}, \vec{\mu})$ that maximizes the long-run average social welfare generated per time unit over an infinite planning horizon.

The following definitions rely on the reader's familiarity with birth-and-death processes (cf. §4.4 of Karlin and Taylor 1997). A policy $(\vec{\lambda}, \vec{\mu})$ is said to be admissible if there exists a unique steady-state distribution $\pi(\vec{\lambda}, \vec{\mu})$ associated with it (given that the system is empty initially) that satisfies the following balance equations:

$$\lambda_n \pi_n(\vec{\lambda}, \vec{\mu}) = \mu_{n+1} \pi_{n+1}(\vec{\lambda}, \vec{\mu}), \quad n = 0, 1, 2, \ldots, \tag{1}$$

and the usual condition

$$\sum_{n=0}^{\infty} \pi_n(\vec{\lambda}, \vec{\mu}) = 1. \tag{2}$$

The long-run average social welfare generated per unit of time under an admissible policy $(\vec{\lambda}, \vec{\mu})$ is given by

$$\gamma_{\vec{\lambda}, \vec{\mu}} = \sum_{n=0}^{\infty} \pi_n(\vec{\lambda}, \vec{\mu})[b(\lambda_n) - c(\mu_n) - h_n]. \tag{3}$$

Next, define

$$\gamma^* \equiv \sup \gamma_{\vec{\lambda}, \vec{\mu}}, \tag{4}$$

where the supremum is taken over all admissible policies $(\vec{\lambda}, \vec{\mu})$. An admissible policy $(\vec{\lambda}, \vec{\mu})$ is said to be optimal if $\gamma_{\vec{\lambda}, \vec{\mu}} = \gamma^*$.

The optimal policy that we derive in the sequel is of the following form $\lambda_n = 0$ for $n \geq N$ (and $\mu_n = M$ for $n \geq N + 1$ by convention), where $N$ is the "optimal buffer size," and it is determined uniquely in the process of deriving our candidate policy. It is crucial to observe that under such a policy we effectively have a system with finite buffer capacity of $N$. Motivated by this, our solution strategy is as follows. First, we restrict attention to policies that effectively truncate the buffer size at $N$ (which will be determined later) and derive a candidate policy by considering the associated system of optimality equations and solving those explicitly. Then, we prove that this candidate policy is indeed optimal for the rate-setting problem in the sense of (4). That is, it is not only optimal among the restricted class of policies, which truncate the buffer size at $N$, but also among the broader class of admissible policies described earlier in this section. It is important to point out that the optimality of the candidate policy that we derive depends critically on our particular choice of the "optimal buffer size" $N$ (cf. conditions (ii) and (iii) of Proposition 1).

### 3.1. A Closely Related Problem Formulation and the Associated Optimality Equation

We now focus attention on the policies that truncate the buffer size at $N$ (which will be determined in §3.2). The standard way of solving such a problem is to consider the associated system of optimality equations, which provides a means for characterizing an optimal policy analytically. Specializing the form of the optimality equations for a semi-Markov decision process with long-run average cost criterion, using the uniformization technique (cf. Bertsekas 1995, p. 267) and assuming by convention that $\mu_0 = 0$ and $\lambda_N = 0$, we arrive at the following set of equations:

$$v_0 = \max_{\lambda\in B}\left\{\frac{b(\lambda)-\gamma}{\Lambda} + \frac{\lambda}{\Lambda}v_1 + \frac{\Lambda-\lambda}{\Lambda}v_0\right\}, \tag{5}$$

$$v_n = \max_{\mu\in A,\,\lambda\in B}\left\{\frac{b(\lambda)-c(\mu)-h_n-\gamma}{\Lambda+M} + \frac{\lambda}{\Lambda+M}v_{n+1} + \frac{\mu}{\Lambda+M}v_{n-1} + \frac{\Lambda+M-\lambda-\mu}{\Lambda+M}v_n\right\}, \quad n=1,\ldots,N-1, \tag{6}$$

$$v_N = \max_{\mu\in A}\left\{\frac{-c(\mu)-h_N-\gamma}{M} + \frac{\mu}{M}v_{N-1} + \frac{M-\mu}{M}v_N\right\}. \tag{7}$$

Here, one interprets $\gamma$ as a guess at the maximum average value (achievable among the restricted class of policies). The vector of unknowns $(v_0, v_1, \ldots, v_N)$ is often called a *relative value function* in average-cost dynamic programming. By rearranging terms, we arrive at the following more compact representation:

$$\gamma = \max_{\lambda\in B}\{b(\lambda) - \lambda(v_0 - v_1)\}, \tag{8}$$

$$\gamma = \max_{\mu\in A,\,\lambda\in B}\{b(\lambda) - \lambda(v_n - v_{n+1}) + \mu(v_{n-1} - v_n) - c(\mu)\} - h_n, \quad n=1,\ldots,N-1, \tag{9}$$

$$\gamma = \max_{\mu\in A}\{\mu(v_{N-1} - v_N) - c(\mu)\} - h_N. \tag{10}$$

It is easy to see that the relative values can only be determined up to an additive constant by (8)–(10), even if $\gamma$ is treated as a known constant. Therefore, it is natural to define the relative value differences

$$y_n = v_{n-1} - v_n, \quad \text{for } n=1,\ldots,N. \tag{11}$$

Then one can reexpress (8)–(10) as follows:

$$\gamma = \max_{\lambda\in B}\{b(\lambda) - \lambda y_1\}, \tag{12}$$

$$\gamma = \max_{\lambda\in B}\{b(\lambda) - \lambda y_{n+1}\} + \max_{\mu\in A}\{\mu y_n - c(\mu)\} - h_n, \quad n=1,\ldots,N-1, \tag{13}$$

$$\gamma = \max_{\mu\in A}\{\mu y_N - c(\mu)\} - h_N. \tag{14}$$

It should be emphasized that the derivation sketched above serves only as motivation in our treatment; the only property of these optimality equations that we require (Proposition 1) will be proved from first principles.

To further reduce the optimality equations, it is natural to define the functions $\phi$ and $\zeta$,

$$\phi(y) = \sup_{\mu\in A}\{y\mu - c(\mu)\}, \quad y \geq 0. \tag{15}$$

For reasons that will become clear in the analysis to follow, we define $\zeta(y)$ only for $y \in [0, b'(0)]$,

$$\zeta(y) = \sup_{\lambda\in B}\{b(\lambda) - y\lambda\}, \quad y \in [0, b'(0)]. \tag{16}$$

Then, the optimality equations (12)–(14) can be rewritten as follows:[6]

$$\gamma = \zeta(y_1), \tag{17}$$

$$\gamma = \zeta(y_{n+1}) + \phi(y_n) - h_n, \quad n=1,\ldots,N-1, \tag{18}$$

$$\gamma = \phi(y_N) - h_N. \tag{19}$$

As observed in George and Harrison (2001, p. 724), given the assumptions on the cost of control $c(\cdot)$ and the set of available service rates that were set forth earlier, it is straightforward to prove the following: First, the supremum in (15) is finite for all $y \geq 0$, and there exists a smallest maximizer $\psi(y) \in A$ that achieves the supremum. Similarly, the supremum in (16) is also finite, and there exists a unique maximizer $\eta(y) \in B$ for $y \in [0, b'(0)]$. Several important properties of these functions will be compiled in §B of the online technical appendix to facilitate our analysis.

We now provide a "verification lemma" allowing us to prove the optimality of a candidate policy derived from a solution of (17)–(19)[7]; its proof is given in §A of the online technical appendix.

Proposition 1. *Given constants $\gamma$ and $N$ and a vector $(y_1, y_2, \ldots, y_N)$, which satisfy*

(i) *$\gamma$ and $(y_1, y_2, \ldots, y_N)$ jointly solve the optimality equations* (17)–(19),

(ii) *$h_n + \gamma \geq \phi(b'(0))$ for $n \geq N+1$,*

(iii) *$0 \leq y_n \leq b'(0)$ for $n = 1, \ldots, N$,*

*let the candidate policy $(\vec{\lambda}^*, \vec{\mu}^*)$ be such that $\lambda_n^* = \eta(y_{n+1})$ for $n = 0, 1, \ldots, N-1$, $\lambda_n^* = 0$ for $n \geq N$, $\mu_n^* = \psi(y_n)$ for $n = 1, \ldots, N$, and $\mu_n = M$ for $n \geq N+1$. If the candidate policy $(\vec{\lambda}^*, \vec{\mu}^*)$ is admissible, then it is also optimal. Moreover, we have that $\gamma_{\vec{\lambda}^*, \vec{\mu}^*} = \gamma = \gamma^*$.*

In the next subsection, we construct a solution to the optimality equation, which satisfies conditions of Proposition 1, and establish the optimality of a candidate policy based on that solution.

[6] We implicitly restrict attention to $y_n \in [0, b'(0)]$.

[7] Conditions (ii) and (iii) of Proposition 1 are auxiliary conditions that determine the optimal buffer size.

### 3.2. Solving the Optimality Equation

In this section, we solve the optimality equations (17)–(19) explicitly. We assume that

$$\phi(b'(0)) > h_1, \tag{20}$$

which assures that our problem is nontrivial. That is, if (20) is not satisfied, then the following analysis shows it is optimal not to accept any customers, which makes the problem uninteresting. Recall that we also assume $b'(0) = \bar{u} < \infty$, which in turn implies that

$$\lim_{n\to\infty} h_n \geq \phi(b'(0)) \tag{21}$$

because $\phi$ is continuous (cf. §B of the online technical appendix) and $\lim_{n\to\infty} h_n = \infty$. The latter assumption derives the conclusion that the system operates as one with a finite buffer under the optimal policy.

The following proposition lies at the heart of our solution method. In particular, it characterizes the optimal buffer size $N$ and the auxiliary functions $y_n(\cdot)$ for $n \leq N$, which will eventually be used to construct a solution to the optimality equations (17)–(19). The following definitions are needed to state Proposition 2. Let $a_1 = 0$ and $b_1 = b(\Lambda)$. Then define

$$y_1(z) = \zeta^{-1}(z) \quad \text{for } z \in [a_1, b_1]. \tag{22}$$

It immediately follows from (22) and Proposition 8 (cf. §B of the online technical appendix) that $y_1(\cdot)$ is continuous and strictly decreasing on $[a_1, b_1]$, with $y_1(a_1) = b'(0)$ and $y_1(b_1) = 0$.

Proposition 2. *There exist $N \geq 1$ pairs $(a_n, b_n)$ and the auxiliary functions $y_n(\cdot)$ for $n = 1, \ldots, N+1$, which are defined inductively, such that the following hold*:

(i) $0 = a_1 < a_2 < \cdots < a_N < a_{N+1} < b_{N+1} < b_N < \cdots < b_2 < b_1 = b(\Lambda)$,

(ii) $h_{n-1} + a_n - \phi(y_{n-1}(a_n)) = 0$ *and* $h_{n-1} + b_n - \phi(y_{n-1}(b_n)) = b(\Lambda)$, $n = 2, \ldots, N+1$,

(iii) $y_n(z) = \zeta^{-1}(z + h_{n-1} - \phi(y_{n-1}(z)))$ *for* $z \in [a_n, b_n]$ *and* $n = 2, \ldots, N+1$,

(iv) $y_n(\cdot)$ *is continuous and strictly decreasing on* $[a_n, b_n]$ *with* $y_n(a_n) = b'(0)$ *and* $y_n(b_n) = 0$ *for* $n = 1, \ldots, N+1$,

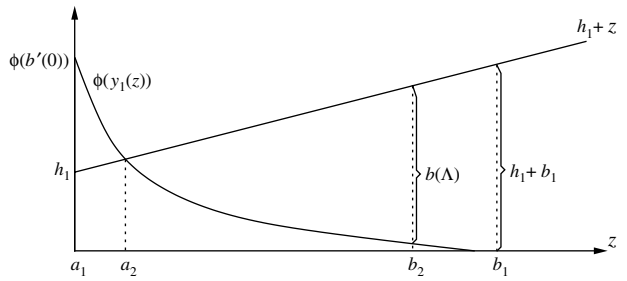
(v) $a_n + h_n < \phi(b'(0))$ *for* $n = 1, \ldots, N$,

(vi) $a_{N+1} + h_{N+1} \geq \phi(b'(0))$.

Proof. We proceed by induction. As the induction basis, we prove that (i)–(v) hold with $N = 1$. To show this, the first step is to construct the pair $(a_2, b_2)$ such that $a_1 < a_2 < b_2 < b_1$ and that $h_1 + a_2 - \phi(y_1(a_2)) = 0$ and $h_1 + b_2 - \phi(y_1(b_2)) = b(\Lambda)$. First, for each $z \in [a_1, b_1]$, we define $f_1(z) = h_1 + z - \phi(y_1(z))$. Clearly, $f_1(\cdot)$ is strictly increasing and continuous on $[a_1, b_1]$, and it follows from (20) that

$$f_1(a_1) = h_1 - \phi(b'(0)) < 0,$$
$$f_1(b_1) = h_1 + b(\Lambda) - \phi(y_1(b_1)) = h_1 + b(\Lambda) > 0.$$

**Figure 1** **An Illustrative Construction of $(a_2, b_2)$**




Therefore, there exists a unique $a_2 \in (a_1, b_1)$ such that $f_1(a_2) = 0$. That is, $h_1 + a_2 - \phi(y_1(a_2)) = 0$. Moreover, because $f_1(a_2) = 0$ and $f_1(b_1) > b(\Lambda)$, there exists a unique $b_2 \in (a_2, b_1)$ such that $f_1(b_2) = b(\Lambda)$. That is, $h_1 + b_2 - \phi(y_1(b_2)) = b(\Lambda)$. Figure 1 illustrates the construction of the pair $(a_2, b_2)$.

It is clear that $f_1(z) = h_1 + z - \phi(y_1(z)) \in [0, b(\Lambda)]$ for $z \in [a_2, b_2]$. Then let

$$y_2(z) = \zeta^{-1}\big(z + h_1 - \phi(y_1(z))\big), \quad z \in [a_2, b_2],$$

which is well defined, continuous, and strictly decreasing with $y_2(a_2) = b'(0)$ and $y_2(b_2) = 0$. Therefore, (i)–(v) hold with $N = 1$.

As our induction hypothesis, suppose that (i)–(v) hold with $N = j - 1$. Then, we have the following two cases to consider:

*Case* 1. $a_j + h_j \geq \phi(b'(0))$. In this case, the inductive definition terminates. We set $N = j - 1$ and $a_n = a_{N+1}$ for $j > N + 1$.

*Case* 2. $a_j + h_j < \phi(b'(0))$. In this case, we proceed with the inductive definition. In particular, the pair $(a_{j+1}, b_{j+1})$ and the function $y_{j+1}(\cdot)$ on $[a_{j+1}, b_{j+1}]$ are constructed as follows: For each $z \in [a_j, b_j]$, we define $f_j(z) = h_j + z - \phi(y_j(z))$. Clearly, $f_j(\cdot)$ is continuous and strictly increasing on $[a_j, b_j]$. Moreover, we have that

$$f_j(a_j) = h_j + a_j - \phi(y_j(a_j)) = h_j + a_j - \phi(b'(0)) < 0,$$
$$f_j(b_j) = h_j + b_j - \phi(y_j(b_j)) > h_{j-1} + b_j \geq b(\Lambda).$$

Therefore, there exists a unique $a_{j+1} \in (a_j, b_j)$ such that $f_j(a_{j+1}) = 0$. That is, $h_j + a_{j+1} - \phi(y_j(a_{j+1})) = 0$. Also, because $f_j(a_{j+1}) = 0$ and $f_j(b_j) > b(\Lambda)$, there exists a unique $b_{j+1} \in (a_{j+1}, b_j)$ such that $f_j(b_{j+1}) = b(\Lambda)$. That is, $h_j + b_{j+1} - \phi(y_j(b_{j+1})) = b(\Lambda)$. Figure 2 illustrates the construction of the pair $(a_{j+1}, b_{j+1})$.

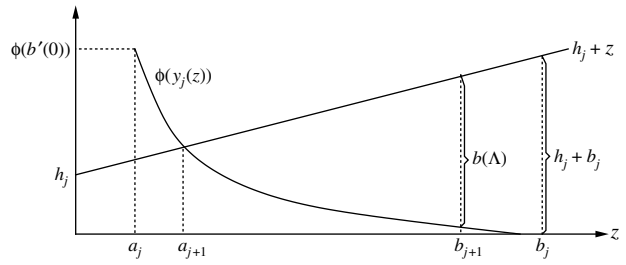
Having constructed the pair $(a_{j+1}, b_{j+1})$, next define

$$y_{j+1}(z) = \zeta^{-1}(z + h_j - \phi(y_j(z))), \quad z \in [a_{j+1}, b_{j+1}]. \tag{23}$$

It is clear that $y_{j+1}(\cdot)$ is continuous and strictly decreasing on $[a_{j+1}, b_{j+1}]$ with $y_{j+1}(a_{j+1}) = b'(0)$ and $y_{j+1}(b_{j+1}) = 0$. Therefore, (i)–(iv) hold with $N = j$.

At this stage, we again have the two cases considered immediately above. That is, we have either

**Figure 2** **An Illustrative Construction of $(a_{j+1}, b_{j+1})$**




$a_{j+1} + h_{j+1} \geq \phi(b'(0))$ or $a_{j+1} + h_{j+1} < \phi(b'(0))$. In the former case, the inductive definition terminates, and in the latter case we proceed with the induction as described immediately above.

It is crucial to observe that $N$ is finite by (21), and $N \geq 1$ by (20). Also, by construction we have that

$$N = \max\{n \geq 1\colon h_n + a_n < \phi(b'(0))\}, \tag{24}$$

and that $h_{N+1} + a_{N+1} \geq \phi(b'(0))$ and, therefore, (i)–(vi) hold. □

In essence, Proposition 2 solves the optimality equations. Using that solution, one can construct an optimal policy. More specifically, Proposition 3 provides a solution to the optimality equations using the auxiliary functions $y_n(\cdot)$ constructed in Proposition 2. Proposition 3 is also useful in establishing that the optimal arrival rate is positive until the system reaches the optimal buffer size, and that the service rate is positive whenever the system is not empty, thereby implying that the truncated system under the optimal policy corresponds to an irreducible Markov chain (cf. Corollary 1).

The following definitions are needed to state Proposition 3, which is proved in §A of the online technical appendix. Let $z^* = a_{N+1}$ and $y_n^* = y_n(z^*)$ for $n = 1, \ldots, N$. Then, it immediately follows from (vi) of Proposition 2 that

$$h_n + z^* \geq \phi(b'(0)) \quad \text{for } n \geq N+1. \tag{25}$$

Proposition 3. *$z^*$ and $(y_1^*, \ldots, y_N^*)$ solve the optimality equations* (17)–(19). *That is,*

$$z^* = \zeta(y_1^*), \tag{26}$$

$$z^* = \zeta(y_{n+1}^*) + \phi(y_n^*) - h_n, \quad n = 1, \ldots, N-1, \tag{27}$$

$$z^* = \phi(y_N^*) - h_N. \tag{28}$$

*Moreover,* $c'(0) < y_1^* < \cdots < y_N^* < b'(0)$.

Having solved the optimality equations, we next propose a candidate policy based on that solution and prove its optimality. In particular, we propose the candidate policy $(\vec{\lambda}^*, \vec{\mu}^*)$, where

$$\lambda_n^* = \begin{cases} \eta(y_{n+1}^*) & \text{if } 0 \leq n \leq N-1, \\ 0 & \text{if } n \geq N, \end{cases} \tag{29}$$

and

$$\mu_n^* = \begin{cases} \psi(y_n^*) & \text{if } 1 \leq n \leq N, \\ M & \text{if } n \geq N+1. \end{cases} \tag{30}$$

The following corollary is an immediate consequence of (29), (30), Proposition 3, (49), and Proposition 5 (cf. §B of the online technical appendix).

Corollary 1. *The service rates of the candidate policy are increasing and the arrival rates are decreasing in the number of customers present in the system. More specifically, we have that*

(i) $0 < \mu_1^* \leq \mu_2^* \leq \cdots \leq \mu_N^* \leq M$;
(ii) $\Lambda \geq \lambda_0^* \geq \lambda_1^* \geq \cdots \geq \lambda_{N-1}^* > 0$.

We now state the main result of §3, which establishes optimality of the candidate policy for the rate-setting problem.

Theorem 1. *The candidate policy $(\vec{\lambda}^*, \vec{\mu}^*)$ defined by* (29) *and* (30) *is optimal for the rate-setting problem in the sense of* (4).

Proof. First, because we assumed that the system is empty initially, it follows from Corollary 1 that there exists a unique steady-state distribution. Thus, the candidate policy is admissible. Also by Proposition 3 and (25), the candidate policy satisfies the necessary conditions so that Proposition 1 can be invoked, which immediately implies that the candidate policy $(\vec{\lambda}^*, \vec{\mu}^*)$ is optimal in the sense of (4). □

## 4. The Price-Setting Problem

In this section, we focus on a system in which the system manager controls the service rates and prices but is unable to directly control arrival rates. The system manager's goal when choosing the service rates and prices is as before: to maximize the long-run average welfare generated. As before, *welfare* is defined as the value gained from service minus cost of capacity minus the delay costs, which correspond to the holding costs in the rate-setting problem. Therefore, the system manager's goal is to coordinate the system to achieve $\gamma^*$ defined in (4). To do that, we construct the optimal state-dependent prices and capacities $(\vec{p}^*, \vec{\mu}^*)$ that the system manager should set to induce the arrival rates $\vec{\lambda}^*$ and service rates $\vec{\mu}^*$ found in Theorem 1.

The prices set by the system manager correspond to the fees customers pay per service request submitted to be processed. These prices are state-dependent. Customers pay the posted price at the time of arrival to the facility, independent of the actual waiting time they experience. Specifically, at each state of the system $n$, the system manager sets both the state-dependent service rate $\mu_n$ and price $p_n$, as well as

posts the true system expected delay. Then, customers, who are rational, utility-maximizing, and delay-sensitive decision makers observe the price for the service and the posted expected delay and decide whether or not to submit their service requests.[8] Clearly, to induce the optimal service rates, the system manager chooses the service rates $\vec{\mu}^*$, as in Theorem 1. It remains to find the optimal prices. To do that, we develop further our model of customer behavior.

In the absence of delays, customers have a positive value from receiving service. We assume that the customers' service values are i.i.d., drawn from a continuous distribution $F$ with p.d.f. $f$, and assumed strictly positive and continuous on $[\underline{u}, \bar{u}]$, where $0 \le \underline{u} < \bar{u} < \infty$. Let $\bar{F} = 1 - F$. If all consumers with value exceeding $u$ submit service requests, the rate at which service requests arrive at the facility is $\lambda = \Lambda \bar{F}(u)$, where $\Lambda$ is the maximum arrival rate. Conversely, the value of the marginal consumer corresponding to arrival rate $\lambda$ is $\bar{F}^{-1}(\lambda/\Lambda)$. Using the notation introduced in §3, $b(\lambda)$ denotes the aggregate (gross) network value rate. Then, the downward-sloping marginal value function $b'(\lambda) = \bar{F}^{-1}(\lambda/\Lambda)$ defines a one-to-one mapping between the arrival rate $\lambda$ and the corresponding marginal value $b'(\lambda)$ with $b'(\lambda) > 0$ and $b''(\lambda) < 0$ for $\lambda < \Lambda$, consistent with the assumptions in §3 (cf. Lippman and Stidham 1977, Mendelson 1985, Dewan and Mendelson 1990, Mendelson and Whang 1990, Afèche and Mendelson 2004).

The customers requesting service are delay-sensitive with a delay of one time unit resulting in $v$ units of customer disutility. The expected delay at state $n$, $D_n$, is a function of the requests' arrival rates $\vec{\lambda}$, the service rates $\vec{\mu}$, and the distribution of service time requirements. We assume that all service requirements are i.i.d., having an exponential distribution with mean 1.

To derive the customer's net utility, who arrives when there are already $n$ customers in the system, from receiving service, we subtract from the customer's gross utility both the price $p_n$ he pays for the service and his expected delay cost $v \cdot D_{n+1}$. Thus, the marginal customer's (expected) net utility when there are $n$ requests already in the queue waiting to be serviced is $b'(\lambda_n) - p_n - v \cdot D_{n+1}$. Customers choose to submit service requests as long as their net utility is positive. Thus, the optimal price at state $n$, given the optimal arrival rates $\vec{\lambda}^*$ and service rates $\vec{\mu}^*$, is

$$p_n^* = b'(\lambda_n^*) - v \cdot D_{n+1}^*, \quad \text{for } n = 0, \dots, N-1, \qquad (31)$$

[8] We make the assumption that delays are posted by the system manager for simplicity. In fact, if customers have full information about the optimal policy and the state of the system on arrival, the expected delay posting is redundant because customers have the information required to calculate the expected delay.

where $D_n^*$ is the expected delay at state $n$ under the optimal policy $(\vec{\lambda}^*, \vec{\mu}^*)$. Thus, using (31), it is straightforward to set the optimal prices, given the optimal arrival rates $\vec{\lambda}^*$, and expected delays $\vec{D}^*$. From Theorem 1, we have the optimal arrival and service rates $\vec{\lambda}^*$ and $\vec{\mu}^*$. It remains to find $\vec{D}^*$.

Denote by $W_i^j$ the expected delay of the $i$th job in the queue under the optimal policy given that the state of the system is $j$ for all $i \le j$; the following proposition characterizes these quantities.

Proposition 4. *Under the optimal policy, which has arrival rates $\vec{\lambda}^*$ and service rates $\vec{\mu}^*$, the expected delay of the $i$th job in the queue, when the state of the system is $j$, is the unique solution to the linear system of equations*

$$W_i^j = \begin{cases} \dfrac{1}{\lambda_j^* + \mu_j^*} + \dfrac{\mu_j^*}{\lambda_j^* + \mu_j^*} W_{i-1}^{j-1} + \dfrac{\lambda_j^*}{\lambda_j^* + \mu_j^*} W_i^{j+1} \\ \qquad \text{for } j < N, \\ \dfrac{1}{\mu_j^*} + W_{i-1}^{j-1} \quad \text{for } j = N, \end{cases}$$

*with $W_0^j = 0$ for all $j$.*

Proof. The recursive equations follow by a first-step analysis. Writing this system of linear equations in the matrix form $AW = B$, where $A$ is an $N(N+1)/2 \times N(N+1)/2$ matrix, and $W$ is the vector $W = [W_1^1, W_1^2, \dots, W_1^N, W_2^2, \dots, W_N^N]$, it is straightforward to see, by inspection, that the matrix $A$ is of full rank. It follows that the system of equations has a unique solution. □

The unique solution to the system of equations in Proposition 4 gives us the expected delays of all customers in the system. To set prices, we need only the expected delay of the arriving customer. That is, $D_n = W_n^n$. Having done that, by (31) we have a full specification of the optimal prices.

To repeat, to maximize welfare in the price-setting problem the system manager first solves the rate-setting problem of §3 with functions $b(\cdot)$, $c(\cdot)$, and the linear holding cost $h_n = \nu n$, corresponding to customers' delay cost. Denoting the resulting optimal policy by $(\vec{\lambda}^*, \vec{\mu}^*)$, the system manager can simply implement the service rate control policy $\vec{\mu}^*$ directly. However, the important difference from the rate-setting problem is that the system manager can control the arrival rate only through the posted prices and expected delays. Proposition 4 characterizes the state-dependent expected delays under the optimal policy for the rate-setting problem, and (31) characterizes the prices that induce those arrival rates. Therefore, the system manager posts these state-dependent prices and expected delays, inducing the optimal arrival rates $\vec{\lambda}^*$.

One may intuitively expect that the optimal price is monotone in the state of the system given that the optimal arrival rate is decreasing in the system state. Equation (31) sheds some light on this issue. In particular, Equation (31) shows that the optimal price is the difference of two increasing functions of the system state. Thus, in general, the price need not be monotone. In the next section, we indeed demonstrate through a numerical study that, although in many cases the price is monotone, it is nonmonotone in some examples. An interesting question is to identify conditions on the model primitives under which the price is necessarily monotone or necessarily nonmonotone. However, the rather implicit characterization of the expected delay $D_n^*$ in Proposition 4 makes such queries analytically intractable.

## 5. A Numerical Example

In §3, we found the solution to the dynamic control problem of arrival and service rates, and in §4 we found the set of prices that induce this optimal system when customers make their individual entry decisions. In this section, we explore numerically the characteristics of the dynamic optimal arrival rates, service rates and prices that our solution induces, and the overall gain in system performance that the dynamic control policy offers.

We assume, for the purpose of the example, that the customers' gross utility from receiving service in the system is a uniformly distributed random variable, that is, the distribution $F$ is uniform over $[\underline{u}, \bar{u}]$. It follows (using the analysis in §4) that the gross utility of the marginal customer from receiving service is given by the linear demand curve $b'(\lambda) = B - 2C\lambda$, where $B = \bar{u}$ and $C = (\bar{u} - \underline{u})/(2\Lambda)$ for all states of the system. Because $B > 2C\Lambda$ and $B, C > 0$ by definition, the value rate function $b(\lambda) = B\lambda - C\lambda^2$ is indeed concave and increasing on $[0, \Lambda]$. We further assume that the capacity cost function is $c(\mu) = \frac{1}{2}\mu^2$, and that the customers' delay cost is $1/m$ per unit of time, for some $m > 0$. From the system manager's point of view, this implies a holding cost of $n/m$ per unit of time, where $n$ is the state of the system.

Integrating the assumptions above with the analysis of §B of the online technical appendix, it is straightforward to arrive at the following closed-form expressions for $\psi(y)$, $\eta(y)$, $\phi(y)$, and $\zeta(y)$:

$$\psi(y) = \begin{cases} y & \text{if } 0 \le y \le M, \\ M & \text{if } M < y; \end{cases}$$

$$\phi(y) = \begin{cases} \frac{1}{2}y^2 & \text{if } 0 \le y \le M, \\ yM - \frac{1}{2}M^2 & \text{if } M < y; \end{cases} \tag{32}$$

$$\eta(y) = \begin{cases} \Lambda & \text{if } 0 \le y \le B - 2C\Lambda, \\ \dfrac{y - B}{-2C} & \text{if } B - 2C\Lambda < y \le B, \\ 0 & \text{if } y > B; \end{cases}$$

$$\zeta(y) = \begin{cases} \Lambda(B - y - C\Lambda) & \text{if } 0 \le y \le B - 2C\Lambda, \\ \dfrac{(B-y)^2}{4C} & \text{if } B - 2C\Lambda < y \le B, \\ 0 & \text{if } y > B. \end{cases} \tag{33}$$

First, we explore the properties of optimal prices. Because the characterization of the optimal prices is rather implicit, we consider special cases to gain insight. If one considers a case where delay costs are small compared with the value rates, it is natural to expect that $p_n \approx b'(\lambda_n^*)$, at least for small $n$. Then, because the optimal arrival rates are increasing and the value rate function $b(\cdot)$ is concave, we intuitively expect the prices to be increasing at least in a range where $n$ is sufficiently small. Indeed, in many examples, this intuition seems to be correct, and the optimal price is monotone increasing in the state of the system; Figure 3 provides an illustrative example of this case. On the other hand, if we consider the "light-traffic" scenario where $\Lambda$ is small, it is natural to expect that $\lambda_n^* = \Lambda$ for most of the system states. In that case, $p_n = b'(\Lambda) - D_{n+1}$, which is decreasing in $n$. This case is illustrated in Figure 4.

Although the intuitive assertions above seem to be correct in some cases, they do not carry over to the general setting (Figure 5 provides a counterexample to both assertions). Therefore, although the overall costs, monetary plus delay cost, experienced by

**Figure 3 An Illustrative Example of Increasing Prices**

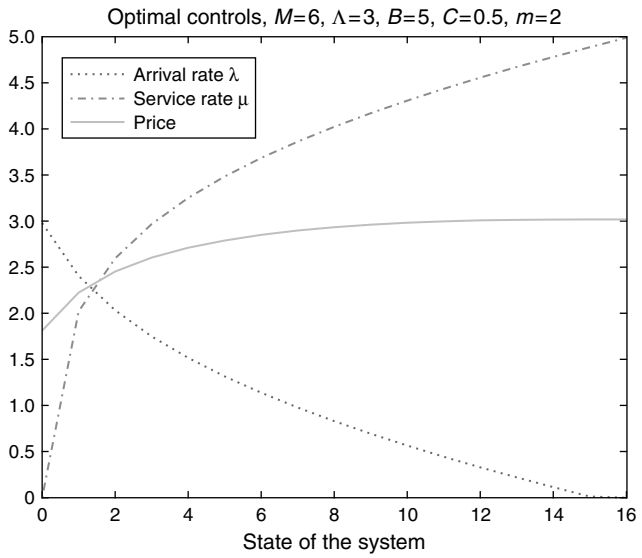
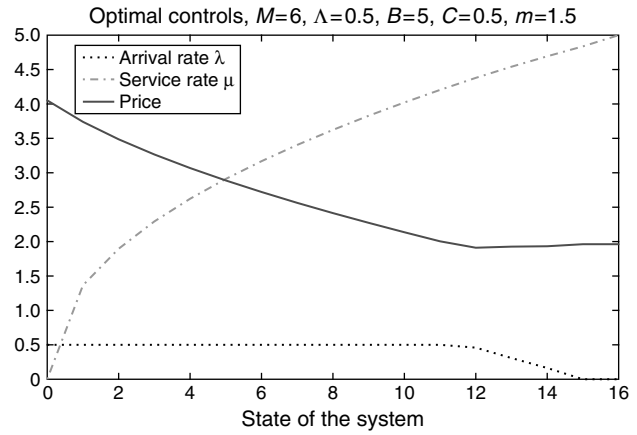


**Figure 4 An Illustrative Example of Decreasing Prices**




customers are increasing with the state of the system, the monetary price charged by the system manager to regulate the system to optimality need not be monotone. Of course, an interesting question is to identify precise conditions on model primitives under which the optimal price is necessarily increasing, decreasing, and nonmonotone. Unfortunately, the rather implicit characterization of the expected delays under the optimal policy makes such precise characterizations analytically intractable.

Next, we fix the parameters $\Lambda = 4$, $M = 6$, $B = 5$, $C = 0.5$ arbitrarily, vary the value of the inverse delay cost $m$, and compare the resulting arrival rates, service rates, and prices of our dynamic control problem to the results of two practically appealing static policies, the M/M/1 and M/M/1/K queues. In both static control problems, the system manager can choose one service rate, one arrival rate, or one price. The customers have no information about the current system state and decide whether or not to submit a service request based on the expected steady-state system delays. The M/M/1 static control problem has been studied by Mendelson (1985) and provides a direct comparison to our dynamically controlled M/M/1 system. However, because our dynamic control policy implicitly implies a finite buffer size, we find it useful to compare our results with that of a statically controlled M/M/1/K system as well. In performing the latter comparison, for each set of system parameters, the system buffer size $K$ is chosen so as to maximize the overall expected system welfare.

**Figure 5 An Illustrative Example of Nonmonotone Prices**

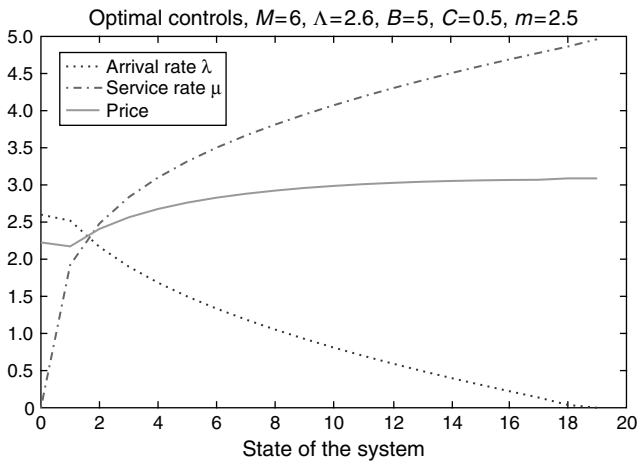
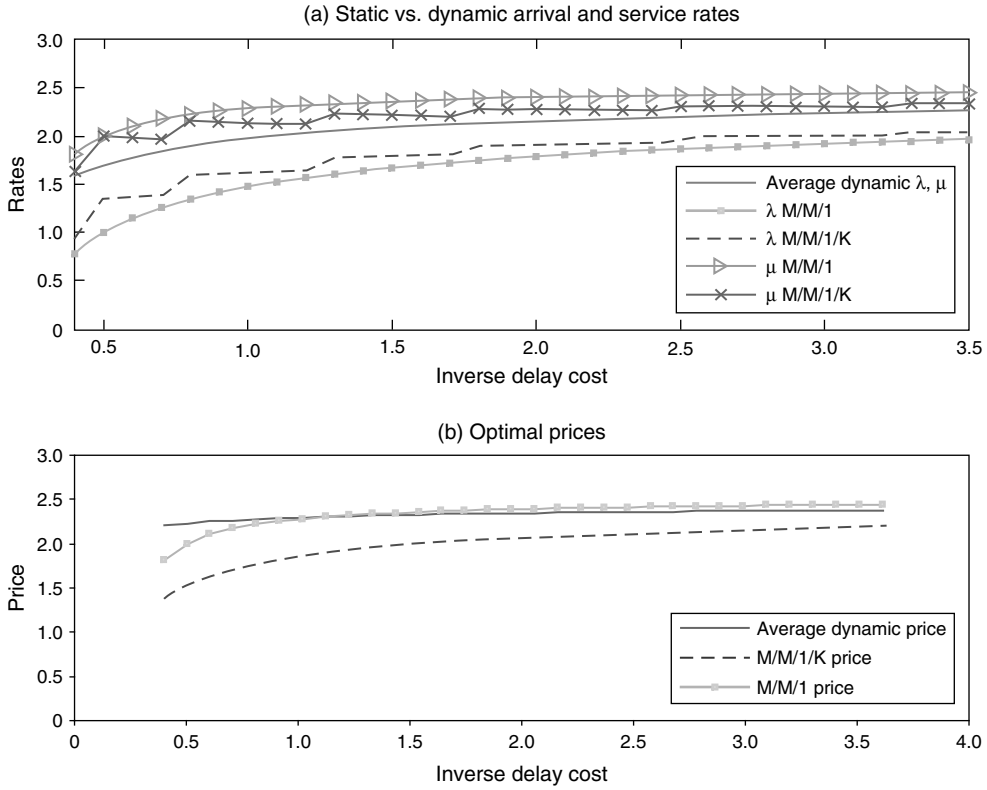



Figure 6(a) illustrates the average optimal arrival rates and service rates in the dynamic problem, where the average is a weighted average with weights set according to the resulting stationary distribution, compared with the static problems' controls. First, note that in the dynamic problem the average service rate $\sum_{n=0}^{N} \pi_i \mu_n$ is always equal to the average arrival rate $\sum_{n=0}^{N} \pi_i \lambda_n$. This follows from the balance equations (1) and the boundary conditions $\mu_0 = 0$ and $\lambda_N = 0$. Thus, not only does the system manager need not worry about stability, in the dynamic setting it is optimal to provide, on average, the service rate that exactly meets the service requirement. This indicates high efficiency attained in the dynamic setting because every unit of service is used to its fullest. Clearly, this cannot be the case in the static M/M/1 system, where we must have $\mu > \lambda$ to guarantee stability. Second, Figure 6(a) illustrates that the service rates are always higher and the arrival rates are always lower in the static systems compared with the average dynamic rates. That is, in addition to using less capacity, more customers are admitted to the system and receive service in the dynamic case. Higher arrival rates imply that customers endowed with relatively low gross utility values are admitted to the system and get to experience its service, whereas these same customers would not be admitted to a static system. As expected, between the two static systems, the controls of the M/M/1/K system are closer to those of the dynamic system. This is because both of these systems have a finite buffer size. Figure 6(b) shows the complementary of this finding in the price-setting problem. The average prices under the dynamic policy are higher than or about the same as prices under the static policies. Combining this with Equation (31) and the fact that average arrival rate is higher for the dynamic policy suggest that delays under the dynamic policy are significantly lower than those under the static policies.

Figure 7(a) illustrates the social welfare increase, in percentages (using a log scale), attained in the optimal dynamic system over the static systems as a function of the inverse delay cost $m$. It is evident, in this example, that the efficiency gain is significant, with the lowest gains being 31.4% and 19.6% over the M/M/1 and M/M/1/K systems, respectively (when $m = 3.5$). The gain in social welfare is increasing with the delay

**Figure 6 Comparison of the Dynamic Rates and Prices with the Static Ones**




*Notes.* Panel (a) illustrates the arrival rates $\lambda$ and service rates $\mu$ in the dynamic and static problems as a function of the inverse demand curve $m$. Panel (b) illustrates a comparison of the prices in both problems, where $\Lambda = 4$, $M = 6$, $B = 5$, $C = 0.5$.

cost. In extreme cases, it might be $\infty$ because it is not worthwhile to operate the system under a static regime (maximum attainable social welfare is zero, which is attained when the system is idle), but operating the system under a dynamic regime induces positive welfare. Note that the M/M/1/K system is more efficient than the M/M/1 system because its finite buffer size serves as an upper bound on the system's maximum experienced delay.

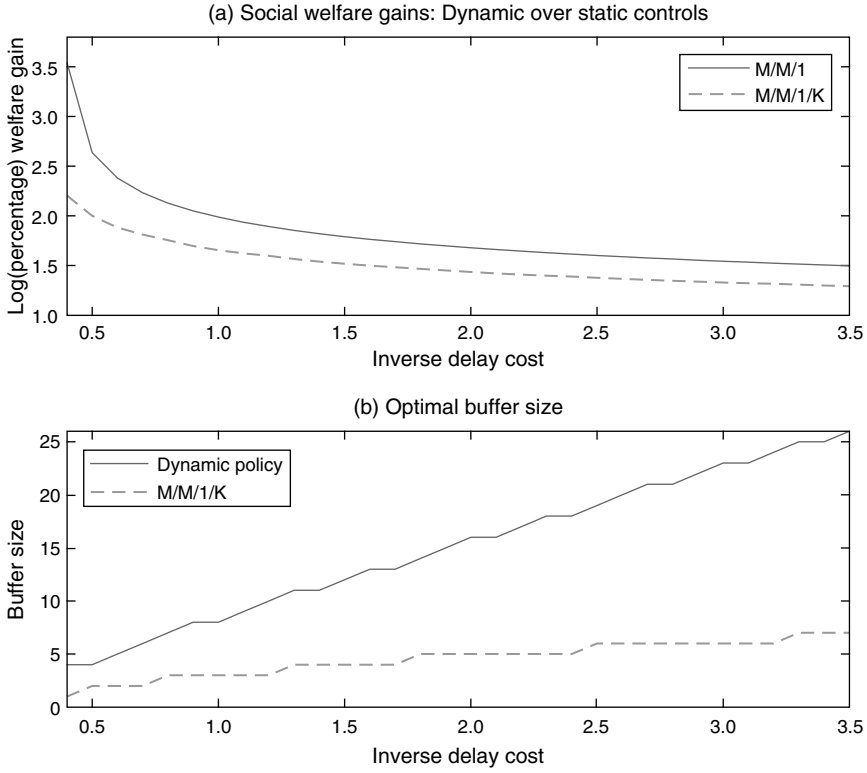
Figure 7(b) illustrates the optimal system buffer size as a function of the inverse delay cost $m$ in both the dynamic and the static M/M/1/K systems. As the delay cost increases, the optimal buffer size decreases, and in our example it varies from $N = 26$ when the delay cost is $v = 1/m = 0.28$ to $N = 3$ when the delay cost is $v = 1/m = 2.5$ in the dynamic system, and from $K = 7$ to $K = 1$ in the M/M/1/K system. The intuitive reasoning is that the buffer size serves as an upper bound on the expected delay in the system. That is, the maximum expected delay, for example, $D_N$ in the dynamic system, is increasing with the buffer size. Therefore, as the cost of delay increases and it is optimal to decrease the expected delays in the system, the optimal buffer sizes are smaller. Because the dynamic system offers more flexibility in controlling the delays than the M/M/1/K system (via increased control of arrival rates and capacities), it has a larger buffer size than the static system.

Combining the insights of Figures 7(a) and (b), we conclude that dynamically controlling a system is most advantageous in systems in which delay costs (or equivalently holding costs) are high and, therefore, optimal buffer sizes are small. Although implementing a dynamic control policy is often a very complex and hard task to manage, our results suggest that this effort is most worthwhile in smaller systems. These are also the systems where the implementation of a dynamic policy is the easiest because the number of possible system states is low.

## 6. Comparative Statics

In most service systems, it is hard to estimate some system parameters reliably, and the parameters may also change over time. Therefore, it is important to understand how the optimal policy and its performance depend on various system parameters. Thus, as a step toward better understanding the nature of the optimal policy and its dependence on the system parameters, we next present its comparative statics. The derivations are provided in §C of the online technical appendix.

**Figure 7 Comparison of the Dynamic Policy with Static Policies**




*Notes.* Panel (a) illustrates the percentage increase in social welfare attained by using the dynamic solution compared with the static solutions, as a function of the inverse delay cost $m$ on a log scale. Panel (b) illustrates the optimal buffer sizes in both the dynamic and M/M/1/K systems, as a function of the inverse delay cost $m$, where $\Lambda = 4$, $M = 6$, $B = 5$, and $C = 0.5$.

In our setting, it is natural to know the cost of capacity function, whereas the delay sensitivity parameter $\nu$ and the value rate function $b(\cdot)$ may not be known precisely. Therefore, the comparative statics are most relevant for the delay-sensitivity parameter $\nu$ and the value rate function $b(\cdot)$. Nevertheless, we also provide the comparative statics for the cost of capacity function $c(\cdot)$, which is assumed to be continuously differentiable and strictly convex in this section, implying that $\phi$ is smooth.

To be specific, we assume that the value rate function is parameterized by $\theta \in (\underline{\theta}, \bar{\theta}) \subset \mathbb{R}$ such that the value rate $b(\lambda, \theta)$ is strictly increasing in $\theta$, and the marginal value rate $\partial b/\partial \lambda$ is nondecreasing in $\theta$, that is,

$$\frac{\partial}{\partial \theta} b(\lambda, \theta) > 0 \quad \text{and} \quad \frac{\partial^2}{\partial \lambda\, \partial \theta} b(\lambda, \theta) \geq 0 \quad \text{for all } \theta,\ \lambda. \tag{34}$$

For instance, taking $b(\lambda) = B\lambda - C\lambda^2$ as in §5, an additive parameterization is given by $b(\lambda, \theta) = (B + \theta)\lambda - C\lambda^2$ for $\theta > 0$, and a multiplicative parameterization is given by $b(\lambda, \theta) = B\theta\lambda - C\lambda^2$ for $\theta > 0$. Both of these parameterizations satisfy our assumptions. Similarly, the cost of capacity function is parameterized by $\alpha \in (\underline{\alpha}, \bar{\alpha}) \subset \mathbb{R}$ such that the cost rate $c(\mu, \alpha)$ is strictly increasing in $\alpha$ and the marginal cost rate $\partial c/\partial \mu$ is nondecreasing in $\alpha$. That is,

$$\frac{\partial}{\partial \alpha} c(\mu, \alpha) > 0 \quad \text{and} \quad \frac{\partial^2}{\partial \mu\, \partial \alpha} c(\mu, \alpha) \geq 0 \quad \text{for all } \alpha,\ \mu. \tag{35}$$

In particular, taking $c(\mu) = \frac{1}{2}\mu^2$ as in §5, for $\alpha > 0$, an additive parameterization is $c(\mu, \alpha) = (\frac{1}{2} + \alpha)\mu^2$, and a multiplicative parameterization is given by $c(\mu, \alpha) = \frac{1}{2}\alpha\mu^2$, both of which satisfy (35).

First, focusing on the linear holding cost case, that is, $h_n = \nu n$ for $n = 0, 1, \ldots,$ we study the dependence of the optimal solution on the delay sensitivity parameter $\nu$. Viewing the optimal solution as a function of $\nu$, it turns out that the optimal value $\gamma^*(\nu)$ is decreasing in $\nu$ as one would expect. To be more specific, defining $L^* = \sum_{n=1}^{N} \pi_n n$ to be the steady-state expected queue length under the optimal policy, one has that

$$\frac{d\gamma^*}{d\nu} = -L^* < 0.$$

We also show that the optimal relative value difference $y_i^*(\nu)$ is strictly increasing in $\nu$ for $i = 1, \ldots, N$. Then, by the monotonicity of $\psi(\cdot)$, it follows that $\mu_i^*(\nu)$ is increasing (strictly increasing unless it is equal

to $M$) for $i = 1, \ldots, N$. Similarly, it follows from the monotonicity of $\eta(\cdot)$ that the optimal arrival rate $\lambda_i^*(\nu)$ is decreasing (strictly decreasing unless it is equal to $\Lambda$) for $i = 0, 1, \ldots, N-1$. Moreover, because the optimal arrival rates are decreasing in $\nu$, we conclude that the optimal buffer size $N$ is also decreasing in $\nu$ as one would expect intuitively.

Second, we study the impact of changes in the value rate function $b(\cdot)$. As one would expect, $\gamma^*(\theta)$ is increasing in $\theta$. In particular, one has that

$$\frac{d\gamma^*}{d\theta} = \sum_{n=0}^{N} \pi_n \frac{\partial}{\partial\theta} b(\lambda_n^*, \theta) = \frac{\partial}{\partial\theta} \mathbb{E}[b(\lambda_n^*, \theta)] > 0.$$

We also show that the optimal relative value difference $y_i^*(\theta)$ is strictly increasing in $\theta$ for $i = 1, \ldots, N$. Then, by the monotonicity of $\psi(\cdot)$, it follows that the optimal service rate $\mu_i^*(\theta)$ is increasing (strictly increasing unless it is equal to $M$) for $i = 1, \ldots, N$. Although one may also expect that the arrival rates are also increasing in $\theta$ for all states, this is not true in general. Indeed, one can easily construct examples where the arrival rate is increasing for small values of the system state and it is decreasing for the large values of the system state; one can also construct examples where the arrival rates are increasing for all states of the system.

Finally, we consider the impact of changes in the cost function. It follows that the optimal value $\gamma^*(\alpha)$ is decreasing as one would expect. More specifically, it turns out that

$$\frac{d\gamma^*}{d\alpha} = -\sum_{n=1}^{N} \pi_n \frac{\partial}{\partial\alpha} c(\mu_n^*, \alpha) = -\frac{\partial}{\partial\alpha} \mathbb{E}[c(\mu_n^*, \alpha)] < 0.$$

We also show that the optimal relative value difference $y_i^*(\alpha)$ is strictly increasing in $\alpha$ for $i = 1, \ldots, N$. Then, by the monotonicity of $\eta(\cdot)$, it follows immediately that the optimal arrival rate $\lambda_i^*(\alpha)$ is decreasing (strictly decreasing unless it is $\Lambda$) for $i = 0, 1, \ldots, N-1$. Because the optimal arrival rates are decreasing in $\alpha$, we also conclude that the optimal buffer size is decreasing in $\alpha$ as one would expect. Although one might intuitively expect that the service rate is decreasing in $\alpha$ for all states, this is not true in general. Indeed, one can construct examples where the service rate is decreasing for large values of the system state and it is increasing for small values of the system state. One can also construct examples where the service rate is decreasing in $\alpha$ for all states of the system.

## 7. Summary and Concluding Remarks

This paper presents the solution to the problem of dynamically controlling the arrival and service rate in a service facility, where the objective is to maximize long-run average system welfare. This solution is then used to solve for the optimal dynamic prices and service rates a system manager should set when serving delay-sensitive, rational customers. Our solution method could be implemented as a tool that supports, or actually manages, decision making in the service facility. For instance, in the context of computing on demand, there could be a server that uses this algorithm to determine service rate (processing power allocation) per task and posted prices dynamically over time.

It turns out that for the system to operate at its optimal arrival and service rates, it needs only to have a finite buffer size, which is characterized by our solution method. The rates themselves are monotone in the state of the system. The arrival rate is decreasing and the service rate is increasing with the number of service requests in the system waiting for service. However, the optimal prices that are set to induce the optimal rates in the price-setting problem need not be monotone. Consequently, even though it may seem intuitive to expect that a system manager who wishes to decrease arrivals would typically raise the price, our analysis shows that this might not be necessary and might lead to too much of a decrease in arrival rate.

Evaluating the increase in social welfare that is achieved by using a dynamic policy instead of the optimal static policies, we find that dynamic policies offer significant welfare gains, and that the higher the customer delay cost is, the more there is to gain from using dynamic policies. In particular, our numerical analysis suggests that dynamic policies are most worthwhile implementing in systems with large delay costs and hence small buffer sizes.

An interesting open problem is to study the case where a system manager strives to maximize her profits (as opposed to maximizing system welfare) in either a monopolistic or competitive market. In principle, for the simpler problem of a monopolist system manager maximizing her profits, one can follow a similar roadmap: First, formulate and solve a rate-setting problem and use that solution to determine the optimal prices. Therefore, it seems that one can focus primarily on the rate-setting problem. In the profit-maximization problem, delay costs incurred by the customers are not part of the system manager's objective. Nevertheless, assuming that the system manager must inform the customers of the true expected delays, the expected delays affect the price (and, hence, the objective to be maximized) necessary to induce a desired arrival rate indirectly through an implicitly defined function, which involves the arrival rates and service rates for all possible states, and that makes the problem intractable. In contrast, for the rate-setting problem studied in §3, the value rate generated for each state $n$ and the associated action

pair $(\lambda_n, \mu_n)$ is an explicit function, namely, $b(\lambda_n) - \nu n - c(\mu_n)$, which makes it amenable to exact analysis. Therefore, it seems that the approach used in this paper cannot be adopted to address the profit-maximization problem, and one may have to resort to an approximate analysis along the lines of Maglaras and Zeevi (2003).

In our analysis, we assume that customers have full information about the state of the system or, equivalently, about their expected delay when making the decision of whether or not to submit a service request. An interesting extension of this work is to examine whether misleading customers about the current system load makes the system manager better off, and whether nontruthful strategies can be sustainable equilibria given that customers have their own beliefs about the system state. Finally, one may also consider extending our model to incorporate strategic customers who may prefer to delay their entrance to the system.

An online supplement to this paper is available on the *Management Science* website (http://mansci.pubs.informs.org/ecompanion.html).

## Acknowledgments

The authors thank Mustafa Akan, Gad Allon, J. Michael Harrison, Sunil Kumar, Haim Mendelson, and the anonymous reviewers for helpful comments. They also thank Konstantinos E. Zachariadis for his technical assistance.

## References

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

Bertsekas, D. P. 1995. *Dynamic Programming and Optimal Control*, Vol. 2. Athena Scientific, Belmont, MA.

Dewan, S., H. Mendelson. 1990. User delay costs and internal pricing for service facility. *Management Sci.* **36**(12) 1502–1517.

George, J. M., J. M. Harrison. 2001. Dynamic control of a queue with adjustable service rate. *Oper. Res.* **49**(5) 720–731.

Hirshleifer, J. 1964. Internal pricing and decentralized decisions. C. Bonini, R. Jaedicke, H. Wagner, eds. *Management Controls: New Directions in Basic Research*. McGraw-Hill, New York.

Karlin, S., H. M. Taylor. 1997. *A First Course in Stochastic Processes*, 2nd ed. Academic Press, New York.

Lippman, S. A., S. Stidham. 1977. Individual versus social optimization in exponential congestion systems. *Oper. Res.* **25**(2) 233–247.

Low, D. W. 1974. Optimal dynamic pricing policies for an M/M/s queue. *Oper. Res.* **22** 545–561.

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

Masuda, Y., S. Whang. 1999. Dynamic pricing for network service: Equilibrium and stability. *Management Sci.* **45**(6) 857–869.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Comm. Assoc. Comput. Machinary* **28** 312–321.

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

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

Paschalidis, I. Ch., Y. Liu. 2002. Pricing in multiservice loss networks: Static pricing, asymptotic optimality, and demand substitution effects. *IEEE/ACM Trans. Networking* **10**(3) 425–438.

Paschalidis, I. Ch., J. N. Tsitsiklis. 2000. Congestion-dependent pricing of network services. *IEEE/ACM Trans. Networking* **8**(2) 171–184.

Stidham, S. 1988. Scheduling, routing, and flow control in stochastic networks. W. Fleming, P. L. Lions, eds. *Stochastic Differential Systems, Stochastic Control Theory and Applications*, Vol. 10. *IMA Volumes in Mathematics and Its Applications*. Springer-Verlag, Berlin, Germany, 529–561.

Stidham, S. 2002. Analysis, design, and control of queueing systems. *Oper. Res.* **50**(1) 197–216.

Talluri, K. T., G. J. Van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Kluwer Academic Publishers, Boston, MA.

Westland, C. J. 1992. Congestion and network externalities in the short run pricing of information system services. *Management Sci.* **38**(7) 992–1009.

Yoon, S., M. E. Lewis. 2004. Optimal pricing and admission control in a queueing system with periodically varying parameters. *Queueing Systems* **47**(3) 177–199.