MANAGEMENT SCIENCE
Vol. 54, No. 6, June 2008, pp. 1132–1146
ISSN 0025-1909 | EISSN 1526-5501 | 08 | 5406 | 1132

informs

DOI 10.1287/mnsc.1070.0836

Dynamic Pricing and Lead-Time Quotation for a Multiclass Make-to-Order Queue

Sabri Çelik
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027, sc2190@columbia.edu

Costis Maglaras
Columbia Business School, Columbia University, New York, New York 10027, c.maglaras@gsb.columbia.edu

This paper considers a profit-maximizing make-to-order manufacturer that offers multiple products to a market of price and delay sensitive users, using a model that captures three aspects of particular interest: first, the joint use of dynamic pricing and lead-time quotation controls to manage demand; second, the presence of a dual sourcing mode that can expedite orders at a cost; and third, the interaction of the aforementioned demand controls with the operational decisions of sequencing and expediting that the firm must employ to optimize revenues and satisfy the quoted lead times. Using an approximating diffusion control problem we derive near-optimal dynamic pricing, lead-time quotation, sequencing, and expediting policies that provide structural insights and lead to practically implementable recommendations. A set of numerical results illustrates the value of joint pricing and lead-time control policies.

Key words: revenue management; dynamic pricing; lead-time quotation; queueing; sequencing; diffusion models

History: Accepted by Wallace J. Hopp, stochastic models and simulation; received June 15, 2005. This paper was with the authors 1 year and 2 weeks for 3 revisions. Published online in *Articles in Advance* March 18, 2008.

1. Introduction

This paper studies the operational and demand control decisions faced by a profit-maximizing make-to-order production firm that offers multiple products to a market of price and delay sensitive customers, emphasizing three features of particular interest: first, the joint use of dynamic pricing and lead-time quotation controls to manage demand; second, the access to a dual sourcing mode that can be used to expedite orders at a cost; and third, the interaction between the demand controls with the operational ones of sequencing and expediting employed by the firm.

Starting with the airline industry, the adoption of *revenue management* strategies has transformed the transportation and hospitality sectors over the past couple of decades, and is now becoming important in retail, telecommunications, entertainment, financial services, health care, and manufacturing. Broadly speaking, these involve the use of sophisticated information technology systems and intense data processing to construct detailed forecasts and quantitative demand models, coupled with the use of differentiated product offerings that are managed through dynamic capacity allocation and/or pricing strategies to maximize the firm's expected profitability. This paper is motivated from manufacturing applications, a notable example of which comes from the automotive industry's effort to produce cars in a make-to-order fashion.[1] In such a setting, the joint consideration of economic and operational decisions makes the manufacturer more responsive to market changes and to fluctuations in its operating environment due to the variability of the demand and production processes. Moreover, the dynamic pricing and lead-time decisions can exploit the customers' heterogeneity in their price and delay sensitivities and drive higher profitability.

In more detail, the production system is modeled as a multiclass $M_q/GI/1$ queue. The system manager can dynamically select its prices and quoted lead times. Operationally, the manager has discretion with respect to the sequencing of orders at the server, and can choose to instantaneously expedite existing orders at a cost. Instantaneous expediting is an idealized model for systems with significant surge capacity, which also enables the firm to satisfy the quoted lead times on all accepted orders. Customers choose what to buy, if any, by trading off price and lead time

[1] For example, BMW claims that 80% of the cars sold in Europe and 30% of those sold in the United States are built to order. When a dealer inputs a potential order to BMW's web ordering service, a target lead time is generated within five seconds. This is typically 11–12 days in Europe and about double that amount in the United States (Edmondson 2003).

against their product valuations and delay preferences. The firm's problem is to select state-dependent pricing and lead-time quotation strategies, and expediting and sequencing policies to maximize its long-run average expected revenue minus expediting cost.

This paper strives to contribute in terms of the modeling and analysis of these problems, as well as the derivation of structural insights that may be of practical value. In terms of modeling, this paper is one of the first to address the joint dynamic pricing and lead-time control problem in a stochastic production environment, and it combines two novel features: first, the incorporation of expediting capability that enriches the model while simplifying the analysis of lead-time guarantees; and second, its treatment of the dynamic lead-time control capability. This is done by committing to offer each "good" at multiple predetermined lead times, and then focusing on pricing these options. Using dynamic pricing the firm can effectively dynamically divert demand across this discrete set of lead-time options. Restricting the possible lead-time options (e.g., one, two, or four weeks) is practical, and simplifies the customer choice model, which reduces to just a function of price. Finally, the customer choice model outlined in §5 builds on extensive marketing research and seems novel.

The resulting problem could be approached using the theory for Markov decision processes, but such a formulation is analytically and numerically intractable. This paper follows the methodology proposed by Harrison (1988) that suggests studying such a model in a regime where its processing resource is almost fully utilized. This leads to a Brownian control problem that is often simpler than the original problem at hand. Apart from this analytical simplification, this operating regime can—at least in some cases—be justified economically (e.g., see Maglaras and Zeevi 2003). Within this framework, the key analytical results of this paper are the following. In §3.1, we propose a diffusion model approximation for the underlying system that is accurate (and asymptotically correct) in settings with high levels of potential demand and processing capacity, respectively, and where the latter is almost fully utilized. We formulate a control problem that is motivated from settings with low capacity costs or production economies of scale, such as in information service networks and some examples of make-to-order manufacturing, respectively. The approximating problem is solved using results by Plambeck et al. (2001) and Ata et al. (2005). Specifically, we use results from Plambeck et al. (2001) to characterize the sequencing and expediting controls that are optimal for the approximating problem (Proposition 1). We then reduce the resulting multidimensional drift control problem to a one-dimensional one in terms of the workload process (Proposition 2), which we solve by adopting results from Ata et al. (2005) (Theorem 1).

The solution of the approximating diffusion control problem leads to intuitive and practically implementable policies for the original problem (see §4), as well as to several structural insights:

(i) Pricing decisions depend on the aggregate system workload and not the product-level queue lengths, and thus tend to vary on a slower time-scale.

(ii) Expediting is done according to a greedy priority rule (from cheapest to most expensive) in order to keep the total workload below a certain level that depends on the predetermined lead-time bounds.

(iii) Sequencing is done according to a dynamic rule that roughly speaking serves the order that is "closest" to violating its lead time.

(iv) In high-volume systems, end-to-end delays are relatively small, and as a result the price differentials that the firm can charge between two variants that only differ in terms of their lead-time guarantees will be small. Our analysis quantified this effect. Extensive numerical results illustrate the value of joint pricing and lead-time control, as well as the performance of the proposed set of policies.

This section concludes with a literature survey. The remainder of this paper is structured as follows. Section 2 describes the problem formulation, §3 proposes and solves its approximating diffusion control problem, the solution of which is interpreted in §4. Section 5 describes a suitable customer choice model. Section 6 summarizes a set of numerical experiments.

1.1. Literature Survey

This paper is related to the literature on dynamic due-date and sequencing control. We refer the reader to Keskinocak and Tayur (2003) for a review of this literature focusing on algorithms for computational solutions to such problems, and to Baker (1984) and Wein (1991) for a review emphasizing the stochastic nature of the production dynamics and its effects on the firm's controls. Duenyas and Hopp (1995) and Duenyas (1995) were the first to incorporate the customer response to the firm's lead-time policy, whereas Keskinocak et al. (2001) and Charnsirisakskul et al. (2006) provide deterministic optimization models for delay and price sensitive demand, respectively.

The second body of research related to our paper focuses on static pricing and sequencing in queues. One stream of this work initiated with Naor (1969) and includes Mendelson (1985) and Mendelson and Whang (1990) studies problems of social welfare optimization for price and delay sensitive customers. Maglaras and Zeevi (2003) established conditions under which revenue maximization in a single-product system induces the heavy-traffic regime. An important partial analog to Mendelson and Whang

(1990) in the context of revenue maximization is the recent paper by Afèche (2004). One insight that follows from the analysis of Afèche (see Maglaras and Zeevi 2004) is that in some cases the firm should create significant lead-time separation between substitutable products in order to increase its revenues. As shown in Maglaras and Zeevi (2004), this effect does not arise in the context of the conventional heavy-traffic conditions that underlie our present work, and, therefore, it is not captured in a natural way in this paper.

Methodologically, our work builds on the work by Mandelbaum and Pats (1995) that derived fluid and diffusion approximations for queues with state-dependent parameters. State-space collapse in Brownian control problems is explained in Harrison and Van Mieghem (1996), and Ata et al. (2005) address a class of diffusion control problems that includes as a special case the one we analyze in §3.3. The formulation of lead-time constraints as upper bounds on the respective queue lengths is from Plambeck et al. (2001) and Maglaras and Van Mieghem (2005), and builds on Reiman's (1984) "snapshot principle." Other related papers are Plambeck (2004), which studied a problem of static pricing and lead-time differentiation for two partially substitutable products, Maglaras (2006), which looked at dynamic pricing and sequencing for a multiproduct queue with price sensitive customers and holding costs incurred by the firm, and Ata (2006), which focused on admission control for a multiclass system with lead-time guarantees and thin arrival streams. Although Ata (2006) does not involve any pricing decisions, its solution builds on Plambeck et al. (2001) and Ata et al. (2005), which is similar to our work. Papers that include expediting or dual-source modes include Plambeck and Ward (2008) and Bradley (2004, 2005). The demand model we propose in §5 borrows from the marketing literature, see, e.g., Bucklin and Gupta (1992); for an overview of demand models for revenue management, see Talluri and van Ryzin (2004).

2. Model Formulation

We consider a make-to-order firm that offers multiple products, indexed by $i = 1, \ldots, I$, to a market of price and delay sensitive customers.

2.1. Lead-Time Guarantees

The firm will offer each "good" at multiple predetermined lead times, whereby a "product" corresponds to a (type of good, lead-time) combination. By dynamically adjusting the product prices, the firm can divert demand from one lead time to another, thus effectively exercising dynamic lead-time control over this predetermined set of options. Specifically, product i orders are quoted a lead-time "guarantee" of d_i time units, which serves as a reliable upper bound for the time it takes from when the order is placed until its production is completed. In a stochastic production setting, such guarantees are typically stated as $\mathbb{P}(\text{delay for a class } i \text{ order} > d_i) \leq \epsilon_i$, where $\epsilon_i \in (0, 1)$ is a desired service level, but the capability for instantaneous expediting allows us to instead impose "hard" lead-time guarantees, i.e., $\epsilon_i = 0$ for all i (explained later).

2.2. Economic Structure and Demand Model

The firm operates in a market with imperfect competition, and has power to influence its vector of demand rates by varying its prices p; $p_i(t)$ denotes the per-unit price for product i at time t. Potential customers arriving at the system at time t observe the current menu of products, which is summarized by the pair $(p(t), d)$ and make their decision of which product to buy, if any. The resulting demand is assumed to be an I-dimensional nonhomogeneous Poisson process with instantaneous rate vector $\lambda(p(t); d)$ determined through a *demand function* that maps a price vector $p \in \mathcal{P}$ into a vector of demand rates $\lambda \in \mathcal{L}(d)$, where $\mathcal{P} \subseteq \mathbb{R}^I$ is the set of feasible price vectors, and $\mathcal{L}(d) = \{x \geq 0\colon\ x = \lambda(p; d),\ p \in \mathcal{P}\} \subseteq \mathbb{R}^I_+$ is the set of achievable demand rate vectors for the lead-time vector $d \in \mathbb{R}^I_+$. Note that $\lambda(\cdot; d)$ only depends on the time t through the price posted at that instance. We assume that $\mathcal{L}(d)$ is a convex set for all $d \in \mathbb{R}^I_+$, and that the demand function $\lambda(p; d)$ is bounded and continuously differentiable in both p and d. If i, j correspond to the same good and $d_i > d_j$, then $p_i < p_j$. In addition, (a) for each product i, $\lambda_i(p; d)$ is strictly decreasing in p_i; (b) for each feasible $p_{-i} = (p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_I)$ and lead-time vector d, there exists a *null* price $p_i^\infty(p_{-i}) \in \mathbb{R}$ such that $\lim_{p_i \to p_i^\infty(p_{-i})} \lambda_i(p_i, p_{-i}; d) = 0$; and (c) the revenue rate $p \cdot \lambda(p; d) = \sum_i p_i \lambda_i(p; d)$ is bounded for all $p \in \mathcal{P}$ and has a finite maximizer. (For any two n vectors, $x \cdot y$ denotes their inner product.)

We will assume that there exists an inverse demand function $p(\lambda; d)$, $p\colon\ \mathcal{L}(d) \to \mathcal{P}$, that maps an achievable vector of demand rates λ into a corresponding vector of prices $p(\lambda; d)$. Although, in general, this inverse mapping need not be unique, it turns out that it is for common examples of demand relations; see Talluri and van Ryzin (2004, §7.3.2). Following a standard practice from revenue management, we shall view the demand rate vector as the firm's control, from which prices can be inferred using the inverse demand function. The expected revenue rate is $r(\lambda; d) := \lambda \cdot p(\lambda; d)$, which is assumed to be bounded, strictly concave, and twice continuously differentiable. Later, we will require the somewhat stronger condition of differentiability along appropriately defined sequences (p^n, d^n) with $d^n \to 0$ (cf. the example in §5). Let $\hat{\lambda}(d) := \arg\max\{r(\lambda; d)\colon\ \lambda \in \mathcal{L}(d)\}$,

and $\Lambda = \sum_i \hat{\lambda}_i(d)$ act as proxy of the total market size for our problem. We will assume that $\hat{\lambda}_i(d) > 0$, i.e., in the absence of any capacity and congestion considerations, the firm would choose to produce all products.

2.3. The System Model

The production facility is modeled as a multiproduct (or multiclass) single-server queue. Orders for each product arrive according to nonhomogeneous Poisson processes, and upon arrival, join dedicated, infinite capacity buffers associated with each product. For each product i the number of orders placed in $[0, t]$ is given by $N_i(\int_0^t \lambda_i(s)\,ds)$, where $N_i(t)$ is a unit rate Poisson process. Service time requirements for product i orders are independent identically distributed (i.i.d.), drawn from some general distribution with mean m_i (rate $\mu_i = 1/m_i$) and finite squared coefficient of variation ξ_i. Let $S_i(t)$ denote the number of class i service completions if the server dedicates t time units in processing class i orders. The processes N_i, S_i are independent of each other and across products. The *load* or *traffic intensity* of the system when the demand vector is λ is defined as $\rho := m \cdot \lambda$.

The firm also controls the order sequencing at the server, and order expediting. Within each product, orders are processed in first-in-first-out (FIFO), the server can only work on one job at any given time, and preemptive-resume type of service is allowed. Under these assumptions, a sequencing policy takes the form of the I-dimensional cumulative allocation process $(T(t)\colon t \geq 0)$ with $T(0) = 0$, where $T_i(t)$ denotes the cumulative time that the server has allocated to class i jobs up to time t. In addition, $T(t)$ is continuous and nondecreasing ($T_i(t)$ increases only when there is at least one job in the queue i or in the server), and satisfies the capacity constraint

$$\sum_i T_i(t) - \sum_i T_i(s) \leq t - s \quad \text{for } 0 \leq s \leq t < \infty. \tag{1}$$

The cumulative idleness process $(I(t)\colon t \geq 0)$ is defined by $I(t) = t - \sum_i T_i(t)$, and is nonnegative, continuous, and nondecreasing. The expediting policy captures actions such as the use of overtime, subcontractors, etc., that increase the firm's short term production capacity, whenever necessary to meet its lead-time guarantees. It is modeled as an I-dimensional process $(B(t)\colon t \geq 0)$ with $B(0) = 0$, where $B_i(t)$ is the cumulative number of product i orders that were expedited in $[0, t]$. We will make the simplifying assumption that expedited orders are produced (and get removed from the corresponding queue) instantaneously. The cost of expediting a class i order is c_i, and without loss of generality we will assume that products are so labeled that $c_1\mu_1 \geq c_2\mu_2 \geq \cdots \geq c_I\mu_I$.

Let $Q_i(t)$ denote the number of product i jobs in the system (i.e., in queue or in service) at time t, that evolves according to (we are assuming for concreteness that $Q(0) = 0$):

$$Q_i(t) = N_i\left(\int_0^t \lambda_i(s)\,ds\right) - S_i(T_i(t)) - B_i(t) \quad \text{for } i = 1, \ldots, I. \tag{2}$$

A control (λ, T, B) will be admissible if in addition to all of the above it is nonanticipating, i.e., decisions at time t only use information that has been made available up to that time.

2.4. Control Problem Formulation

The profit-maximization problem for the stochastic queueing model is the following: choose admissible demand, sequencing and expediting policies (λ, T, B), respectively, to maximize the long-run average expected profit given by[2]

$$\lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\int_0^t r(\lambda(s); d)\,ds - c \cdot B(t)\right]. \tag{3}$$

2.5. Discussion of Modeling Assumptions

The Poisson nature of the demand processes is needed to be able to justify the diffusion models used in §3. The assumption regarding the general service time distributions is innocuous because the diffusion analysis only uses its first and second moments. As in most papers on pricing in queues and revenue management, our model assumes that self-interested customers decide whether to place an order based solely on the price (and lead-time) vector at the time of their arrival; i.e., they are strategic in making purchase selections by explicitly or implicitly optimizing some form of a personal utility function, but not strategic in selecting the timing of their arrival in response to the firm's pricing strategy. This reduces the firm's pricing problem to one of optimal intensity control not involving a game-theoretic analysis; see Lariviere and Van Mieghem (2004) for a discussion of this point and a justification of the Poisson arrival process assumption as the equilibrium of such a game for a related model. Expediting control capability is a common business practice that fits naturally in our problem formulation, and it allows the firm to be able to satisfy all outstanding lead-time guarantees. In the context of the asymptotic analysis of this paper, this could also be achieved by ejecting orders from a queue, exercising "hard" admission control (i.e., turn off a demand stream), or simply through a more "aggressive" pricing policy.

[2] The formulation in (3) is derived using a standard result for intensity control (see Brémaud 1980, §II.2) from the primitive problem: choose p, T, B to maximize $\lim_{t\uparrow\infty}(1/t)\mathbb{E}[\int_0^t p(s) \cdot dA(s) - c \cdot B(t)]$, where $A_i(t) = N_i(\int_0^t \lambda_i(s)\,ds)$.

3. An Approximating Diffusion Control Problem and Its Solution

This section studies a diffusion model approximation for the problem described in §2. In more detail, §3.1 develops the approximating diffusion control problem, §3.2 reduces this multidimensional formulation to a one-dimensional problem in terms of the aggregate system workload, which is solved in closed-form in §3.3. The solution is interpreted into an implementable policy in §4.

3.1. Heuristic Derivation of an Approximating Diffusion Control Problem

The approximating diffusion control problem derived below is summarized through conditions (6)–(10) for the system dynamics and the control objective (13). A rigorous justification of this model through an asymptotic analysis would follow the line of argument found in Plambeck et al. (2001) with two modifications to account for the dynamic pricing capability of our model that makes the demand rate state dependent, and the fact that orders are not blocked but are expedited.

3.1.1. Background on the Large-Scale Behavior of the Single-Server Queue. Consider a single-class $M/M/1$ queue with arrival rate Λ, service rate $\mu > \Lambda$, and traffic intensity $\rho = \Lambda/\mu < 1$. Denote by $\mathbb{E}Q$ and $\mathbb{E}W$ the expected queue length and waiting time that jobs spend in the system, respectively. Elementary results from queueing theory give that $\mathbb{E}Q = \rho/(1-\rho)$ and $\mathbb{E}W = \rho/(\Lambda(1-\rho))$. The primary motivation for the work in this paper is high-volume stochastic production systems, which in the context of the stylized $M/M/1$ queue described above would entail that both Λ, μ are large. The production economies of scale that are inherent in the above expressions imply that unless the load ρ is close to one, the steady state behavior of the system is one where the queue lengths are modest and the expected waiting time experienced by consumers is negligible. With that in mind, it is natural to study these systems in heavily loaded regimes, where, more precisely, $\rho = 1 - \theta/\sqrt{\Lambda}$ for some constant $\theta > 0$. This is the so-called heavy-traffic operating regime, with Λ acting as a proxy for the "size" of the system. It can be shown to be economically optimal for an $M/M/1$ production queue that offers one product to a market of price and delay sensitive consumers.[3] For this regime the above expressions imply that the queue length and the resulting waiting time are of order $\sqrt{\Lambda}$ and $1/\sqrt{\Lambda}$, respectively; specifically, $\mathbb{E}Q = \sqrt{\Lambda}/\theta - 1$ and $\mathbb{E}W = 1/(\theta\sqrt{\Lambda}) - 1/\Lambda$. Each arriving order waits for roughly $\sqrt{\Lambda}$ jobs to be processed before itself commences service, but the overall level of congestion is moderate due to the production economies of scale of large scale systems embodied in the shorter service times (of order $1/\Lambda$). These results suggest that typical lead times in such systems will also be of order $1/\sqrt{\Lambda}$; shorter lead times will be impossible to cope with in steady-state, and longer ones will be trivially satisfied.

3.1.2. Approximating Diffusion Model. The above insights hold for more general models that include multiple products and networks of servers; see Reiman (1984, 1988), Plambeck et al. (2001), Harrison (2003). The diffusion model proposed below is for a multiproduct, single-server system, whose drift parameters θ_i are state dependent. It could be justified as an asymptotic limit in settings where the market sizes for each product (one measure of which are the quantities given by $\hat{\lambda}(d)$) and the processing rate vector μ grow proportionally large, and where the lead-time bounds grow large relative to the processing time requirement of each order, but remain modest in absolute terms. Recall the definitions of $\hat{\lambda}(d)$, Λ and define

$$\bar{\lambda}(d) := \arg\max\left\{ r(\lambda; d):\ \sum_i \lambda_i/\mu_i = 1,\ \lambda \in \mathcal{L}(d) \right\} \quad (4)$$

to be the demand rate vector that maximizes the instantaneous revenue rate subject to the constraint that the server is fully utilized.[4] As we scale the market potential and processing rate for each product in a way that $\hat{\lambda}_i(d)$ grows large, $\bar{\lambda}_i(d)$ defined via (4) will also grow large. We will assume that $\bar{\lambda}_i(d) > 0$ for all i, i.e., that it is optimal to produce all products in this deterministic planning problem, and let $\bar{\rho}_i = \bar{\lambda}_i(d)/\mu_i$. This assumption will ensure that all product variants will be offered in the approximating model that we will propose in the sequel, which, in broad terms, will be used to characterize the demand rate fluctuations around the nominal demand vector $\bar{\lambda}(d)$.

The candidate controls: Any candidate dynamic drift control $\lambda(t)$ can be expressed as $\lambda(t) = [\bar{\lambda}(d) - \nu(t)]^+$ for some choice of $\nu(t)$. Motivated by the preceding discussion on the $M/M/1$ queue, we will rewrite $\nu(t)$ as $\sqrt{\Lambda}\theta(t)$ and consider drift controls of the form

$$\lambda(t) = [\bar{\lambda}(d) - \sqrt{\Lambda}\theta(t)]^+. \quad (5)$$

The traffic intensity at time t is given by $\rho(t) = [1 - \sum_i (\sqrt{\Lambda}/\mu_i)\theta_i(t)]^+$. If Λ, μ are proportionally large,

[3] Maglaras and Zeevi (2003) established this result for an $M/M/N$ system model, and a simplified version of their argument could handle the $M/M/1$ queue discussed here. The key underlying assumptions are that delay costs are linear and additive, and that the demand function is elastic.

[4] We assume that this problem is feasible for the choice of lead-time vector d under consideration, i.e., that there exists a vector of nonnegative prices (including $p = 0$) for which the resulting demand will utilize all the capacity.

then $(\sqrt{\Lambda}/\mu_i)$ is of order $1/\sqrt{\Lambda}$, and the system will operate in the desired heavy-traffic regime. Since $\bar{\lambda}(d)$ is also of order Λ, if θ is of moderate magnitude, then $\lambda(t) > 0$ for all t.

The sequencing decisions are expressed by the cumulative server allocation process $T(t)$. For each product i, define $V_i(t) = \bar{\rho}_i t - T_i(t)$ to measure the deviation between the cumulative time allocated into processing class i orders up to time t and the "nominal" service requirement for that product predicted by the utilization vector $\bar{\rho}$. The intuition here is that if Λ, μ are large and the server is almost fully utilized, then the cumulative time that the server has allocated in processing class i orders must be close to $\bar{\rho}_i t$. Note that the cumulative idleness up to time t is $I(t) = \sum_i V_i(t)$.

System dynamics: In large scale systems the cumulative arrivals and service completions up to any time t can be approximated using the Strong Approximation Theorem for the associated stochastic processes. We use Ethier and Kurtz (1986, Corollary 7.5.5), which allows us to approximate a process $Y(t)$ with a given mean and variance by $Y(t) = X(t) + o(\sqrt{t})$ almost surely (a.s.), where $X(t)$ is a Brownian motion with the same mean and variance with $Y(t)$, and the notation $f(x) = o(g(x))$ denotes that $f(x)/g(x) \to 0$ as $x \uparrow \infty$, Applying this result to the service completion process $S_i(T_i(t))$ and using the fact that $T_i(t) = \bar{\rho}_i t - V_i(t)$ we get that for all i, $S_i(T_i(t)) = \mu_i \bar{\rho}_i t - \mu_i V_i(t) + X'_{s,i}(\mu_i \bar{\rho}_i t - \mu_i V_i(t)) + o(\sqrt{\mu_i t})$, a.s., where $X'_{s,i}$ are independent, standard Brownian motions. Recall that μ is of same order as Λ. Motivated by the preceding discussion for the $M/M/1$ queue in heavy traffic we proceed by assuming optimistically that $\mu_i V_i(t)$ is itself of order $\sqrt{\mu_i t}$, which gives that

$$S_i(T_i(t)) = \mu_i \bar{\rho}_i t - \mu_i V_i(t) + \sqrt{\Lambda}\sigma_{s,i} X_{s,i}(t + \bar{\varphi}(t)) + o(\sqrt{\Lambda}) \quad \text{a.s.},$$

where $\bar{\varphi}(t)$ is of order $1/\sqrt{\Lambda}$ and $X_{s,i}$ are independent, standard Brownian motions, $\sigma^2_{s,i} = \tilde{\lambda}_i \xi_i$, where $\tilde{\lambda} := \bar{\lambda}(d)/\Lambda$. Similarly, $N_i(\int_0^t \lambda_i(s)\,ds) = \int_0^t \lambda_i(s)\,ds + X'_{a,i}(\int_0^t \lambda_i(s)\,ds) + o(\sqrt{\Lambda})$ a.s., where $X'_{a,i}$ are independent, standard Brownian motions. Expanding the expression for $\lambda_i(t)$ and assuming that $\lambda_i(t) > 0$ for all t (this is expected to hold for large enough Λ), we get that

$$N_i\left(\int_0^t \lambda_i(s)\,ds\right) = \bar{\lambda}_i(d)t - \sqrt{\Lambda}\int_0^t \theta_i(s)\,ds + \sqrt{\Lambda}\sigma_{a,i} X_{a,i}(t + \tilde{\varphi}(t)) + o(\sqrt{\Lambda}),$$

where $\tilde{\varphi}(t)$ is of order $1/\sqrt{\Lambda}$, $X_{a,i}$ are independent standard Brownian motions, and $\sigma^2_{a,i} = \tilde{\lambda}_i$. Plugging into (2), dividing by $\sqrt{\Lambda}$ and defining

$$Z(t) = Q(t)/\sqrt{\Lambda}, \quad Y(t) = \sqrt{\Lambda}V(t), \quad \text{and} \quad D(t) = B(t)/\sqrt{\Lambda},$$

as motivated by the scaling relations briefly highlighted for the $M/M/1$ queue, we get that

$$Z(t) = Z(0) - \int_0^t \theta(s)\,ds + MY(t) + \Sigma X(t) - D(t) + o(1),$$

where X is an I-dimensional standard Brownian motion with $X(0) = 0$ a.s. on some filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}; \mathcal{F}_t, t \geq 0)$, with $(\mathcal{F}_t, t \geq 0)$ being the filtration generated by X, $M = \text{diag}(\tilde{\mu}_1, \ldots, \tilde{\mu}_I)$, $\Sigma^2 = \text{diag}(\sigma_1^2, \ldots, \sigma_I^2)$, and where $\tilde{\mu}_i = \mu_i/\Lambda$ and $\sigma_i^2 = \sigma^2_{a,i}(1 + \xi_i)$ respectively, and the notation $\text{diag}(x)$ denotes a diagonal matrix with entries $x_1, x_2, \ldots, x_I$. This heuristic argument suggests the following approximating diffusion model:

$$dZ(t) = -\theta(t)dt + \Sigma dX(t) + M dY(t) - dD(t), \quad Z(0) = 0, \tag{6}$$

$$L(t) = \sum_i Y_i(t), \quad L(\cdot) \text{ is continuous, nondecreasing with } L(0) = 0, \tag{7}$$

$$D(\cdot) \text{ is continuous, nondecreasing with } D(0) = 0, \tag{8}$$

$$Z(t) \geq 0, \quad \forall\, t \geq 0 \quad \text{and} \quad Y, D \text{ are nonanticipating with respect to } X. \tag{9}$$

In the above model, Z represents the queue length process, D is the expediting policy, Y is the allocation control (measuring deviations from the nominal allocation), and L represents the scaled cumulative idleness. The control processes Y, D are right continuous with left-hand limits (RCLL). As in Plambeck et al. (2001), the lead-time constraints take the form

$$Z(t) \leq b \quad \text{for } t \geq 0, \quad \text{where } b_i = \tilde{\lambda}_i \tilde{d}_i \quad \forall\, i, \tag{10}$$

where $\tilde{d} = d\sqrt{\Lambda}$. The latter can be interpreted as follows. Since (10) indicates that $Q_i(t) \lesssim \bar{\lambda}_i(d) d_i$ and $\bar{\lambda}_i(d) d_i$ is roughly the number of class i arrivals in the last d_i time units, this constraint implies that the waiting times for orders in queue i at time t is less than or equal to d_i, as required.

3.1.3. Performance Criterion and the Resulting Diffusion Control Problem. The expediting costs in $[0, t]$ are given by $c \cdot B(t) = \tilde{c} \cdot D(t)$ for $\tilde{c} = c\sqrt{\Lambda}$. To express the revenue term of the objective function when Λ, μ are large, we first define

$$\kappa := \frac{\hat{\lambda}(d) - \bar{\lambda}(d)}{\sqrt{\Lambda}}. \tag{11}$$

Using (11) one can rewrite $\lambda(t)$ as $\lambda(t) = [\hat{\lambda}(d) - \sqrt{\Lambda} \cdot (\kappa + \theta(t))]^+$. Recall that $\tilde{d} = d\sqrt{\Lambda}$ and let $\tilde{r}(x; \tilde{d}) = r(\Lambda x; \tilde{d}/\sqrt{\Lambda})/\Lambda$ be the revenue function normalized

again by Λ, which acts as proxy for the system size.[5] Let $\hat{x} = \arg\max \tilde{r}(x; \tilde{d}) = \hat{\lambda}(d)/\Lambda$ be the corresponding maximizer. The large-scale behavior of the single-server queue suggests that the system can operate at almost full resource utilization while jobs experience only modest delays, which, in turn, implies that the firm needs to apply only moderate demand adjustments through $\theta(t)$, such that $\sqrt{\Lambda}\theta(t) \ll \bar{\lambda}(d)$. This allows us to approximate the revenue function using a Taylor expansion as follows:

$$\begin{aligned} r(\lambda(t); d) &= \Lambda \tilde{r}\big([\hat{x} - (\kappa + \theta(t))/\sqrt{\Lambda}]^+; \tilde{d}\big) \\ &= \Lambda \tilde{r}(\hat{x}; d) - [\kappa + \theta(t)] \cdot A[\kappa + \theta(t)] + o(1), \end{aligned}$$

where

$$A = -(1/2)\nabla^2 \tilde{r}(\hat{x}; \tilde{d}). \tag{12}$$

The second expression for $r(\lambda(t); d)$ uses the observation that for large Λ, μ, the demand vector is strictly positive, and then applies a Taylor expansion of $\tilde{r}(\cdot; \tilde{d})$ around $\hat{x}$; the first order term is missing because $\nabla \tilde{r}(\hat{x}; \tilde{d}) = 0$ by the optimality of $\hat{x}$.

In the sequel, to minimize technical complexity and comply with the restrictions imposed in Ata et al. (2005) in solving diffusion control problems similar to ours, we will assume that θ is bounded by a large constant K. Using the definitions of κ, A, $\tilde{c}$ given above, and restricting attention to Markovian, stationary, bounded drift controls, the preceding analysis suggests the following diffusion control problem: choose a nonanticipating measurable drift function $\theta(t) \in [-K, K]^I$ for all $t \geq 0$, and nonanticipating, RCLL, allocation and expediting policies Y and D to minimize

$$\limsup_{t\to\infty} \frac{1}{t} \mathbb{E}\left[\int_0^t \{2\kappa \cdot A\theta(s) + \theta(s) \cdot A\theta(s)\}\, ds + \tilde{c} \cdot D(t)\right] \tag{13}$$

subject to (6)–(10).

3.1.4. Remark on (11). The vector κ measures the difference between the demand rate vectors that maximize the instantaneous revenue rates with and without the capacity constraint, $\bar{\lambda}(d)$ and $\hat{\lambda}(d)$, respectively. Given a set of system parameters, it is, of course, always possible to define κ as in (11), even if the difference $\hat{\lambda}(d) - \bar{\lambda}(d)$ is significant, but the implicit assumption imposed through the definition of κ is that it is meaningful to express this difference as a second-order term that is proportional to $\sqrt{\Lambda}$, even though both $\bar{\lambda}(d)$ and $\hat{\lambda}(d)$ are themselves of order Λ. This says that the capacity rate vector is close to the nominal processing requirement implied by the vector of (capacity unconstrained) revenue maximizing demand rates. Our numerical results will illustrate that the performance of the heuristics derived from our diffusion approximation is best, and the overall value of dynamic pricing is highest, when κ is moderate. This is reminiscent of other papers in revenue management, such as Gallego and van Ryzin (1994), which showed (a) the central role played by the demand rates given by $\hat{\lambda}$ and $\bar{\lambda}$ that either maximize revenues or optimally deplete capacity by a specified time, and (b) that the effect of tactical dynamic pricing adjustments is more significant for load factors close to one. Finally, we note that the normalization of the expediting cost coefficient from c to $\tilde{c}$ makes the expediting costs of the same order of magnitude as the revenue corrections due to dynamic pricing, leading into a control problem formulation that captures the trade-off between these two elements.

3.2. Reduction to the Equivalent Workload Formulation

The first step in analyzing (6)–(10) and (13) establishes that the optimal pair of allocation and expediting policies (Y, D) derived in Plambeck et al. (2001) is also optimal for our problem that incorporates dynamic drift rate control capability. (The model analyzed in Plambeck et al. (2001) involved admission rather than expediting decisions, but the two are analytically equivalent.) The optimal (Y, D) yield a one-dimensional equivalent workload formulation for our problem (see Harrison and Van Mieghem 1996 for background), which will be used to derive the optimal dynamic drift control $\theta(\cdot)$.

We start with some background material. The system workload process is defined by $W(t) := \widetilde{m} \cdot Z(t)$, where $\widetilde{m}_i = 1/\tilde{\mu}_i$. The workload dynamics are given by

$$dW(t) = -\widetilde{m} \cdot \theta(t)dt + \sigma_w dX_w(t) + dL(t) - dU(t), \tag{14}$$

$$U(t) = \widetilde{m} \cdot D(t), \quad U(\cdot) \text{ is continuous and nondecreasing, } U(0) = 0, \tag{15}$$

where L satisfies (7), D satisfies (8), and $X_w(t)$ is a standard Brownian motion with infinitesimal variance $\sigma_w^2 = \sum_i (1 + \xi_i)\widetilde{m}_i^2 \tilde{\lambda}_i$. Note that $Z(t) \leq b$ implies that

$$W(t) \in [0, \widetilde{w}] \quad t \geq 0 \quad \text{for } \widetilde{w} := \widetilde{m} \cdot b. \tag{16}$$

The next result specializes some of the results of Plambeck et al. (2001) to our model. Specifically, we

[5] To understand the proposed scaling it is easiest to think of the demand function in the form $N \times F(p, d)$, where N is the number of potential customers per-unit time and $F_i(p, d)$ is the fraction of these customers that would choose product option i given the price and delay menu (p, d). The function $F(\cdot, \cdot)$ encodes the heterogenous customer valuations and delay preferences. The approximate model we study in this paper would correspond to a setting where N grows large but the mapping F, and as a result the customer preferences, stay unchanged. That is, customers would choose which product to purchase if any based on the price vector p and the delay vector $\tilde{d}/\sqrt{\Lambda}$.

show that it is optimal to (i) only expedite orders of the cheapest class I when the workload $W(t)$ reaches its upper bound $\widetilde{w}$, and (ii) schedule orders according to the "least slack policy" that maintains the relative backlogs defined as $\eta_i(t) := (Z_i(t)/b_i)$ equal for all i, by giving priority to the class that is closest to violating its lead-time bound. (All proofs are relegated to the online appendix, which is provided in the e-companion.)[6]

Proposition 1. *Fix any admissible drift* $(\theta(t),\ t \geq 0)$ *and consider the problem of minimizing*

$$\limsup_{t\to\infty} \frac{1}{t}\mathbb{E}[\tilde{c}\cdot D(t)], \tag{17}$$

by choosing (Y, D) *subject to the constraints* (6)–(10), *with* $Z(0) = 0$, *and under the labeling assumption that* $\tilde{c}_1\tilde{\mu}_1 \geq \tilde{c}_2\tilde{\mu}_2 \geq \cdots \geq \tilde{c}_I\tilde{\mu}_I$. *Let* L, U *be the pair of continuous, nondecreasing processes, with* $L(0) = U(0) = 0$ *such that*

$$\int_0^t 1_{\{W(s)>0\}}\, dL(s) = 0 \quad \text{and} \quad \int_0^t 1_{\{W(s)<\widetilde{w}\}}\, dU(s) = 0 \qquad t \geq 0, \tag{18}$$

where $W(t)$ *satisfies condition* (14) *and the constraint* (16). *Then, the policy*

$$D_i(t) = 0 \quad \forall\, i \neq I, \qquad D_I(t) = \tilde{\mu}_I U(t), \tag{19}$$

and Y *given in* (EC.1 of the online appendix)[7] *is optimal. In addition, under this policy* $Z(t) = (b/\widetilde{w})W(t)$ *for all* $t \geq 0$.

Under the control (Y, D) identified in (EC.1) and (19),

$$Z(t) = \frac{b}{\widetilde{w}}W(t), \tag{20}$$

and the problem specified by (6)–(10) and (13) reduces to one of selecting the drift control θ to minimize

$$\limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\int_0^t \{2\kappa\cdot A\theta(s) + \theta(s)\cdot A\theta(s)\}\, ds + \tilde{c}_I\tilde{\mu}_I U(t)\right] \tag{21}$$

subject to (14), (16), and conditions (18) that uniquely identify the processes L, U (see Harrison 1985, §2.4). Expressions (EC.1), (19), and (20) can then be used to specify Y, D, Z, respectively.

This problem can be further simplified by noting that the system dynamics in (14) are only affected by the drift control θ through its aggregate value $\psi := \widetilde{m}\cdot\theta$. This will allow us to reformulate the above problem in terms of the one-dimensional control $\psi \in [-K', K']$. To that end, let

$$\begin{aligned}\theta^* = \tilde{f}(\psi) &:= \arg\min\{2\kappa\cdot A\theta + \theta\cdot A\theta\colon\ \widetilde{m}\cdot\theta = \psi\} \\ &= \frac{A^{-1}\widetilde{m}}{\widetilde{m}\cdot A^{-1}\widetilde{m}}(\psi + \widetilde{m}\cdot\kappa) - \kappa,\end{aligned} \tag{22}$$

and note that the revenue loss at θ^* is

$$2\kappa\cdot A\theta^* + \theta^*\cdot A\theta^* = \frac{(\psi + \widetilde{m}\cdot\kappa)^2}{\widetilde{m}\cdot A^{-1}\widetilde{m}} + \kappa\cdot A\kappa. \tag{23}$$

Using these definitions and removing the constant term $\kappa\cdot A\kappa$ of (23) from the objective function, we can formulate the following diffusion control problem:

Proposition 2. *The diffusion control problem* (6)–(10) *and* (13) *is equivalent to the following formulation*: *choose a nonanticipating measurable function* $\psi(t) \in [-K', K']$ *for* $t \geq 0$ *to minimize*

$$\limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\int_0^t \left\{\frac{[\psi(s) + \widetilde{m}\cdot\kappa]^2}{\widetilde{m}\cdot A^{-1}\widetilde{m}}\right\} ds + \tilde{c}_I\tilde{\mu}_I U(t)\right]; \tag{24}$$

subject to

$$dW(t) = -\psi(t)dt + \sigma_w dX_w(t) + dL(t) - dU(t), \tag{25}$$

(16) *and* (18). *Specifically, for every feasible control* ψ *there exists a feasible control* (Y, D, θ) *for* (6)–(10) *with the same performance*; *and, for every feasible control* (Y', D', θ) *for* (6)–(10), *there exists a feasible control* ψ *with at least as good performance*.

3.3. Solution of the Equivalent Workload Formulation

The equivalent workload formulation is a one-dimensional drift control problem for a diffusion that is constrained to lie in the interval $[0, \widetilde{w}]$. We will solve the problem described immediately above using results derived in Ata et al. (2005), for which we need to restrict attention to Markovian, stationary, and bounded controls. To avoid introducing new notation we "overload" the use of the symbol ψ to now be a function of the workload position, i.e., $\psi(W(t))$, instead of being a function of time $\psi(t)$ as in the previous subsection. Let $\alpha_w = (\widetilde{m}\cdot A^{-1}\widetilde{m})^{-1} > 0$ and $\kappa_w = \widetilde{m}\cdot\kappa$, which together with the Markovian structure of the drift controls allows us to rewrite the objective as

$$\limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\int_0^t \{\alpha_w[\psi(W(s)) + \kappa_w]^2\}\, ds + \tilde{c}_I\tilde{\mu}_I U(t)\right]. \tag{26}$$

Let

$$h(\tilde{c}_I\tilde{\mu}_I, \alpha_w, \kappa_w, \widetilde{w}, \sigma_w) := \tilde{c}_I\tilde{\mu}_I - \left[2\alpha_w\kappa_w - \left(\frac{\widetilde{w}}{2\alpha_w\sigma_w^2} + \frac{1}{2\alpha_w\kappa_w}\right)^{-1}\right].$$

[6] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

[7] For completeness, (EC.1) states that $Y_i(t) = -\widetilde{m}_i[Z(0) + \Theta(t) + \Sigma X(t) - D(t)] + (\widetilde{m}_i b_i/\widetilde{w})W(t)$, for $i = 1, \ldots, I$ and $\forall\, t \geq 0$.

THEOREM 1. *Consider the problem of selecting a nonanticipating measurable function* ψ: $[0, \widetilde{w}] \to [-K', K']$ *to minimize* (26) *subject to* (16), (18), *and* (25). *Then, if* $h(\tilde{c}_I\tilde{\mu}_I, \alpha_w, \kappa_w, \widetilde{w}, \sigma_w) = 0$, *the optimal workload drift rate* $\psi^*(w)$ *is*

$$\psi^*(w) = -\left(\frac{w}{\sigma_w^2} + \frac{1}{\kappa_w}\right)^{-1}, \tag{27}$$

if $h(\tilde{c}_I\tilde{\mu}_I, \alpha_w, \kappa_w, \widetilde{w}, \sigma_w) > 0$,

$$\psi^*(w) = \sqrt{\frac{\zeta_1}{\alpha_w}} \tan\left[\frac{w}{\sigma_w^2}\sqrt{\frac{\zeta_1}{\alpha_w}} - \arctan\left(\kappa_w\sqrt{\frac{\alpha_w}{\zeta_1}}\right)\right], \tag{28}$$

where ζ_1 *is the unique positive solution of* (EC.7 of the online appendix), *and otherwise if* $h(\tilde{c}_I\tilde{\mu}_I, \alpha_w, \kappa_w, \widetilde{w}, \sigma_w) < 0$,

$$\psi^*(w) = \sqrt{\frac{\zeta_2}{\alpha_w}} - 2\sqrt{\frac{\zeta_2}{\alpha_w}}\left[1 - \exp\left\{-\frac{2w}{\sigma^2}\sqrt{\frac{\zeta_2}{\alpha_w}} + C\right\}\right]^{-1}, \tag{29}$$

where $C = \ln((\sqrt{\zeta_2/\alpha_w} - \kappa_w)/(\sqrt{\zeta_2/\alpha_w} + \kappa_w))$ *and* ζ_2 *is the unique solution of* (EC.10 of the online appendix) *that lies in* $(0, \alpha_w\kappa_w^2)$. *The optimal product-level drift rate is given by*

$$\theta^*(w) = \tilde{f}(\psi^*(w)) = \alpha_w A^{-1}\widetilde{m}(\psi^*(w) + \widetilde{m}\cdot\kappa) - \kappa. \tag{30}$$

REMARK. Note that ψ^* is monotonically increasing in w in all three cases above, which, in turn, implies that $\theta^*(w)$ is increasing in w, as expected in light of (5). Expression (28) corresponds to the case where the expediting cost $\tilde{c}_I\tilde{\mu}_I$ is high, whereas expression (29) is for the case where expediting is cheap. Since the workload is bounded by $\widetilde{w}$, one can always select the constant K' in the theorem so that the boundedness constraint is never binding, and therefore not restrictive.

3.4. Optimal Static Pricing Solution

The numerical experiments in §6 will contrast the proposed solution against one that uses static pricing. This amounts to selecting the constant vector $\theta \in \mathbb{R}^I$ to minimize

$$\limsup_{t\to\infty} \frac{1}{t}\left[\mathbb{E}\int_0^t \{2\kappa\cdot A\theta + \theta\cdot A\theta\}\,ds + \tilde{c}\cdot D(t)\right] \tag{31}$$

subject to (6)–(10). Using Propositions 1 and 2, this is reduced to the problem

$$\min_{\psi\in\mathbb{R}}\left\{\alpha_w[\psi + \kappa_w]^2 + \tilde{c}_I\tilde{\mu}_I \limsup_{t\to\infty}\mathbb{E}\left[\frac{1}{t}\,U(t)\right]\right\} \tag{32}$$

subject to (16), (18), and (25). That is, the optimal allocation and expediting policies Y, D are the same with those for the dynamic drift control problem. The workload process evolves like a Brownian motion with infinitesimal drift ψ and infinitesimal variance σ_w^2 in the interval $[0, \widetilde{w}]$, with exponential steady-state distribution with mean $\sigma_w^2/(2\psi)$. This leads to the optimization problem

$$\min_{\psi\in\mathbb{R}}\left\{\alpha_w[\psi + \kappa_w]^2 + \frac{\tilde{c}_I\tilde{\mu}_I\psi}{e^{2\psi\widetilde{w}/\sigma_w^2} - 1}\right\}, \tag{33}$$

where the expression for the second term above is given in Harrison (1985, pp. 88–90).

THEOREM 2. *Consider the problem of selecting a constant vector* $\theta \in \mathbb{R}^I$ *to minimize* (33) *subject to* (6)–(10). *Let* ψ^* *be the minimizer of* (33). *The optimal drift vector is* $\theta^* = \alpha_w A^{-1}\widetilde{m}(\psi^* + \widetilde{m}\cdot\kappa) - \kappa$.

4. The Proposed Solution

This section interprets the optimal diffusion controls derived above into an implementable set of pricing, expediting and sequencing policies for the original problem posed in §2. For completeness, we also recapitulate the definitions of the various parameters used in computing the drift function ψ that plays an important role in the pricing policy. Recall the definitions of $\bar{\lambda}(d)$, $\hat{\lambda}(d)$, and Λ and let $\tilde{\lambda} = \bar{\lambda}(d)/\Lambda$, $\tilde{\mu} = \mu/\Lambda(\widetilde{m} = 1/\tilde{\mu})$, $\tilde{c} = c\sqrt{\Lambda}$, $\tilde{d} = d\sqrt{\Lambda}$, and $\widetilde{w} = \sum_i \widetilde{m}_i\tilde{\lambda}_i\tilde{d}_i$. Finally, given κ and A as defined in (11) and (12), let $\kappa_w = \widetilde{m}\cdot\kappa$ and $\alpha_w = (\widetilde{m}\cdot A^{-1}\widetilde{m})^{-1}$.

Pricing: Our analysis shows that the optimal demand control in the approximating diffusion model is a function of the aggregate workload in the system $W(t) = m\cdot Q(t)$. Specifically, given the workload position w, the manager computes the target resource utilization $\rho^*(w)$ as

$$\rho^*(w) := [1 - (1/\sqrt{\Lambda})\cdot\psi^*(w\sqrt{\Lambda})]^+, \tag{34}$$

where $\psi^*(\cdot)$ is the monotonically increasing function specified in Theorem 1, and then selects the demand rate vector

$$\lambda^*(w; d) = \arg\max\{r(\lambda; d):\ \lambda\cdot m = \rho^*(w), \lambda\in\mathcal{L}(d)\}. \tag{35}$$

The corresponding pricing strategy can be inferred via the inverse demand relation $p(\lambda; d)$.

Sequencing: Priority is given to class

$$i^* = \arg\max_i \frac{Q_i(t)}{b_i'}, \quad \text{where } b_i' = \bar{\lambda}_i(d)d_i - \delta_i,$$

and $\delta_i \in \mathbb{R}$ is a "tunable" parameter discussed below. This is the "least relative slack" policy of Plambeck et al. (2001).

Expediting: Orders are expedited when the total workload reaches $W' = m\cdot b'$ according to a rule that gives priority to class I then class $I-1$ (if no class I orders are in queue) and so on. This is the simplest interpretation of the diffusion policy derived in §3.2 in

the context of the original problem at hand, keeping in mind that products are labeled in a way that $c_1\mu_1 \geq c_2\mu_2 \geq \cdots \geq c_I\mu_I$. Perhaps a more intuitive policy would take into account the relative age of different classes by expediting class I orders whenever $Q_i(t) \geq b'_i$ for some class i. This policy "corrects" for the possibility that some class may be violating its lead-time bound while the workload is below its threshold W'.

Another interpretation of the diffusion control expediting policy is in terms of the age of the jobs in the system. That is, the system expedites when the workload reaches the upper bound $\tilde{w}$, which under the proposed sequencing rule corresponds to the instances (in the diffusion model) where the queue lengths and the corresponding waiting times are about to violate their upper bounds. Hence, one could interpret the expediting policy as one that keeps track of the age of jobs in the system and expedite accordingly; this is harder to implement in terms of information requirements, but may still be executable in some production environments.

We conclude this section with a few comments on the structure of the proposed policies.

(i) *Lead-time constraint formulation, sequencing and the choice of the tunable parameter* δ: The parameter b'_i serves as a proxy for the number of class i arrivals in d_i time units, and thus maintaining the queue lengths below their respective thresholds would tend to imply that the corresponding lead-time guarantees are met with high probability; cf., Maglaras and Van Mieghem (2005), Plambeck et al. (2001).

The threshold b'_i is derived from (10), appropriately adjusted through some "tunable" parameters δ_i to correct for the modeling idealizations of the diffusion model, and the state-dependent nature of the demand rate. Specifically, while if in the diffusion model the queue length is less than or equal to its respective threshold, then the lead times of all orders in queue are met with probability one; in the original system this may only be true with high probability. By decreasing the original threshold b_i by $\delta_i > 0$, the manager adds some safety margin in her calculation to guard against this issue. The second effect that factors in this calculation is that because $\lambda^*_i(w; d)$ is decreasing in the workload w, $b_i = \bar{\lambda}_i(d)d_i$ is in fact an overestimate of the expected number of arrivals in d_i time units. Hence, it is likely that when a queue i reaches b_i the age of the oldest orders in queue will have already violated their lead-time bound. Again, a lowering of the threshold b_i will adjust for that effect as well. Table 1 in §6 offers some insight on its effect on the probabilities of expediting and lead-time violation. The magnitude of these two effects is small, and so is the size of the parameter δ, which can be selected via simulation. Note that both of these effects vanish asymptotically, and thus the diffusion model cannot be used to compute these δs.[8] To further simplify this calculation, we note that, instead of computing all the δ_is, one could find one tunable parameter ω that adjusts the workload threshold to $W' = (\sum_i m_i\bar{\lambda}_i(d)d_i) - \omega$, and then set δ_is equal to $\omega\mu_i/(\sum_i m_i)$. This simpler procedure exploits the behavior of the optimally controlled system, and was suggested in a recent paper by Rubino and Ata (2008).

(ii) *Workload dependence and time-scale of price changes*: The proposed sequencing policy tries to distribute the workload in fixed proportions across the various queues, therefore making $W(t)$ an accurate proxy for the system state; this equivalence is exact in the diffusion model. It therefore suffices to restrict attention to pricing and expediting policies that are functions of the workload. This simplifies analysis and has an important implication on the relative time scale for these decisions. Specifically, known results for queues operating in heavy traffic predict that order interarrival and service times are much shorter than the typical queueing times encountered in the system, which in turn are much shorter than the time required for the workload (and the respective queue lengths) to experience significant fluctuations. Pricing changes and expediting decisions occur on the slowest time scale on which the system workload evolves, which is practically appealing. As an example, orders may be arriving every 30 minutes, queueing delays and lead times may be of order of a week, and the workload (and the prices) may fluctuate on a monthly basis.

(iii) *Lead-time control*: The dynamic lead-time control decisions are effectively captured in the demand control $\lambda^*(W(t); d)$ that specifies how to optimally divert demand from one lead-time class to another. This will become clearer in §5, where we study in more detail a particular demand model that is suitable for the problem under consideration.

5. A Choice Model for Joint Pricing and Lead-Time Control

The customer choice model we propose below satisfies the assumptions imposed in §2 and seems suitable for the choice problem considered in this paper. It builds on a framework that has been used extensively in the marketing literature (see Bucklin and Gupta 1992), which postulates that customers make their purchase selection in two stages: first, they decide which product category to buy from, if at all; and second, they choose a specific product from their

[8] Similar small safety parameters that are selected by trial-and-error are common in policies extracted via a fluid or diffusion model analysis. Computing these parameters is fairly simple after one has specified the pricing, sequencing and expediting policies, which was the hard part of the analysis.

selected category—these are referred to as "purchase incidence" and "product or brand choice," respectively (see Bucklin and Gupta (1992) for a discussion of such models and their practical use). From our viewpoint, this model captures the substitution effects among otherwise identical products offered at different price and lead-time combinations, while maintaining analytical and numerical tractability and being suitable for calibration using real data as indicated by the associated voluminous marketing literature.

As explained in §2, a product corresponds to a (type of good, lead-time) combination. To simplify the exposition we first describe the choice model for the case of one good offered at multiple lead times, and then extend it to consider many goods offered at multiple lead times each.

5.1. One Good Offered at Multiple (Price, Lead-Time) Combinations

Following Bucklin and Gupta (1992), we model the probability that a customer will buy one of the products (the "purchase incidence") using a binary logit function

$$\mathbb{P}(\text{inc}) = \frac{e^{V(p,d)}}{1+e^{V(p,d)}},$$

$$\text{where } V(p,d) = \gamma_0 + \gamma_1 \log\left(\sum_i e^{-b_1 p_i - b_2 d_i}\right). \quad (36)$$

$V(p,d)$ corresponds to the deterministic component of the purchase utility from all offered products specified through p, d. The constants $\gamma_0, \gamma_1, b_1, b_2$ are meant to be calibrated from observed data (see Bucklin and Gupta 1992). This purchase incidence probability is equivalent to saying that each arriving customer assigns a utility $V(p,d)+\epsilon$ to the offered group of products, where ϵ is an i.i.d. random component that differentiates potential customers, and follows a logistic distribution with shape parameter equal to one; the effect of different shape parameters can be rolled into $\gamma_0, \gamma_1, b_1, b_2$.

Each arriving customer also has a random delay sensitivity parameter χ for the offered good, which is assumed to be drawn from a continuous distribution with finite support, is independent of ϵ and i.i.d. across customers. Given that an arriving customer decides to purchase a product, she makes her selection to minimize her cost given by $p_i + \chi d_i$, i.e.,

$$\mathbb{P}(i \mid \text{inc}) = \mathbb{P}(p_i + \chi d_i \le p_j + \chi d_j,\ \forall\, j \ne i); \quad (37)$$

the form of (37) captures the price-delay trade-off faced by each customer, and differs from the multinomial logit model used in Bucklin and Gupta (1992). Assuming that potential customers arrive according to a Poisson process with rate Λ_o, products are labeled in such a way that $d_1 < d_2 < \cdots < d_I$, and that prices are ordered in reverse, i.e., $p_1 \ge p_2 \ge \cdots \ge p_I$, we get that

$$\begin{aligned}\lambda_i(p;d) &= \Lambda_o \cdot \mathbb{P}(\text{inc}) \cdot \mathbb{P}(i \mid \text{inc}) \\ &= \Lambda_o \cdot \frac{e^{V(p,d)}}{1+e^{V(p,d)}} \\ &\quad \cdot \mathbb{P}\left(\max_{j>i} \frac{p_i - p_j}{d_j - d_i} \le \chi \le \min_{k<i} \frac{p_k - p_i}{d_i - d_k}\right).\end{aligned}$$

5.2. Many Goods Offered at Multiple (Price, Lead-Time) Combinations

The above model can be extended to allow for many goods offered at potentially multiple (price, lead-time) combinations by incorporating the decision of which good to purchase in the incidence probability, leaving unchanged the second decision stage where a customer selects which product option of a particular good to purchase. Specifically, suppose that there are K goods, with $K < I$, and let $\mathscr{C}(k)$ be the set of products that correspond to good k. Let

$$V^k(p,d) = \gamma_0^k + \gamma_1^k \log\left(\sum_{i \in \mathscr{C}(k)} e^{-b_1^k p_i - b_2^k d_i}\right),$$

denote the purchase utility from good k products, and the constants γ_0^k, γ_i^k, b_1^k, b_2^k are meant to have been calibrated from observed data. Assume that a customer's net purchase utility for good k products is $V^k(p,d)+\epsilon^k$, where the ϵ^ks are i.i.d. across goods, Gumbell distributed random variables with shape parameter one. The incidence probability, which now reduces to the decision of which good to purchase, if any, is computed using the multinomial logit model (Talluri and van Ryzin 2004, §7.2):

$$\mathbb{P}(\text{select good } k) = \frac{e^{V^k(p,d)}}{1+\sum_j e^{V^j(p,d)}}.$$

The product choice is done according to (37), specialized only to products in the set $\mathscr{C}(k)$.

5.3. A Structural Property of This Demand Model and Its Impact on Lead-Time Control

An important step in implementing the policy described in §4 is the computation of the optimal demand vector given a target aggregate traffic intensity $\rho^* = 1 - \psi^*/\sqrt{\Lambda}$; see §4 and Theorem 1. Simple but long algebraic manipulations show that in the parameter regime of interest in this paper, i.e., where Λ and μ are large and $d = \tilde{d}/\sqrt{\Lambda}$ for some $\tilde{d} > 0$, the solution to the problem for the single good case

$$\max_p \left\{\sum_i p_i \lambda_i(p;d):\ \sum_i \lambda_i(p;d) = \mu(1-\psi/\sqrt{\Lambda})\right\},$$

is of the form

$$p_i = \bar{p} + \frac{\pi_i}{\sqrt{\Lambda}} + \frac{z(\psi)}{\sqrt{\Lambda}} + o(1/\sqrt{\Lambda}), \quad (38)$$

where $\bar{p}$ is common across all products and is independent of Λ, and π_i, $z(\psi) \in \mathbb{R}$; the state-dependent price correction $z(\psi)$ is also common across all products. The observation that all prices can be expressed as small perturbations around a common price $\bar{p}$ follows from the fact that the various products correspond to the same good offered at slightly different lead-time guarantees (recall that demand for each $d = \tilde{d}/\sqrt{\Lambda}$), and therefore they have to be priced similarly. A brief sketch of the justification of (38) for the case of two products is given in the online appendix. A similar result can be proved for multiple goods offered at different price, lead-time combinations, in which case the terms $z^k(\psi)$ will depend on the good k. Note that using a pricing policy of the form given in (38) together with the lead-time bounds $d/\sqrt{\Lambda}$ remains well defined even as we let Λ grow large and d decrease to zero.

This property is a consequence of the demand model considered in this section and the assumption that Λ is large, and does not depend on any aspect of the diffusion model itself, including the implicit assumption embodied in (11). The resulting form of the pricing policy in (38) has an important implication on the firm's "lead-time control policy." Specifically, policies of the form given in (38) imply that $(p_i - p_j) = (\pi_i - \pi_j)/\sqrt{\Lambda}$ for all ψ, which plugging into (37) gives that $\mathbb{P}(i \mid \text{inc})$ is independent of ψ! That is, the firm adjusts its nominal price level through $z(\psi)/\sqrt{\Lambda}$ to modulate the aggregate order volume placed with the system, while keeping the fractions of the total order flow that choose each lead-time option constant. That is, it does not choose to divert demand from one lead time to another as the system gets congested, but rather scales down the demand for all products by a common factor by adjusting its price to affect the incidence probability.

5.4. Comments on Modeling Price and Delay Sensitive Demand

An alternative to the two-stage decision process of our model would consider customers arrive with a random valuation v and delay sensitivity parameter χ, and make their purchase decisions in one stage according to

$$\lambda_i(p;d) = \Lambda_o \mathbb{P}(v - p_i - \chi_i d_i \geq 0,\ p_i + \chi d_i \leq p_j + \chi d_j\ \forall j \neq i).$$

Although this may appear to be a more direct and natural model of demand, it is harder to analyze because evaluating the above expression involves the joint distribution of (v, χ). Moreover, its complexity increases significantly when one considers multiple goods offered in different (price, lead-time) options, where customers arrive with different valuations for each of these goods. The hierarchical decision approach of our model assumes that problem away by restricting attention to a structure where the probability that a customer selects a particular service is given by the product of two probabilities, where one depends on the valuation and the other on the delay sensitivity. Apart from its inherent tractability, extensive studies reported in the marketing literature indicate its versatility in capturing customer demand in diverse and complicated settings.

6. Numerical Results and Concluding Remarks

This section reports on a set of numerical experiments that illustrate the effectiveness of dynamic over static pricing, as well as the impact of lead-time control. The latter is achieved by offering the same good at two lead-time options whose prices are dynamically adjusted. We compare the effectiveness of the proposed heuristics against the solution of an approximate Markov decision process (MDP) formulation. We use the following notations. $\mathbb{E}[\pi]$: expected profit; $\mathbb{E}[\rho]$: expected load; $\mathbb{P}(LT)$: probability of violating the lead-time constraint; $\mathbb{P}(exp)$: probability of expediting; $\overline{TR}$: average tardiness. Average tardiness is computed conditional on job being late.

The MDP formulation that we used in our experiments is only an approximation of the original problem at hand, where the lead-time constraint has been replaced by upper bounds on the respective queue lengths, and where expediting occurs whenever these bounds are reached. In contrast, an exact MDP formulation would have to expand the state descriptor to keep track the (continuous) age of each job in every queue, increasing significantly the analytical and numerical complexity of the resulting problem. This MDP approximation is, in part, motivated by the analysis of §3, as well as the use of a related MDP formulation in Ata (2006), where he compared the performance of an admission control heuristic in a system with lead-time guarantees with the solution of such an approximate MDP formulation. In more detail, the MDP formulation adopted in the sequel imposes upper bounds on the queue lengths given by $K_i := \bar{\lambda}_i(d)d_i - \delta_i$, where the δs are tunable parameters adjusted to make sure that the probability of lead-time violation is no greater than 3%; cf. discussion following Table 1. In the single lead-time case (Tables 2 and 3) the resulting optimization problem amounts to selecting the demand rates as a function of the queue length position q, for all $q \leq K$ to maximize the long-run average rate of profits for an $M/M/1/K$ system that expedites orders at a cost of $\$c$ when $q = K$. The problems with two lead-time options of Table 4 involve the analysis of a two-dimensional Markov chain, where in addition to the two state-dependent demand rates the system manager also selects the sequencing vector at each state.

Table 1 The Effect of the Expediting "Tune"-Parameter δ

	Dynamic					Static				
δ	$\mathbb{E}[\pi]$	$\mathbb{E}[\rho]$	$\mathbb{P}(exp)$	$\mathbb{P}(LT)$	$\overline{TR}$	$\mathbb{E}[\pi]$	$\mathbb{E}[\rho]$	$\mathbb{P}(exp)$	$\mathbb{P}(LT)$	$\overline{TR}$
0	20.66	0.98	0.021	0.097	0.57	19.24	0.93	0.060	0.0041	0.29
1	20.40	0.97	0.030	0.070	0.48	18.99	0.92	0.068	0.0018	0.27
2	20.10	0.97	0.039	0.046	0.43	18.78	0.92	0.074	0.0005	0.22
3	19.83	0.97	0.046	0.027	0.38	18.53	0.91	0.080	0.0002	0.21
4	19.43	0.97	0.058	0.014	0.34	18.20	0.91	0.090	0.0000	0.10

Note. Typical standard deviations for the various estimated performance measures were of the order of 0.1% of the estimated parameter value.

Table 2 The Effect of Capacity Imbalance (κ) or Load Factor ($\hat{\rho}$)

			MDP			Dynamic		Static	
μ	$\hat{\rho}(d)$	κ	$\mathbb{E}[\pi]$	$\mathbb{E}[\rho]$	$\mathbb{P}(exp)$	Gap (%)	$\mathbb{P}(exp)$	Gap (%)	$\mathbb{P}(exp)$
4	1.25	0.45	20.83	0.94	0.003	4.80	0.046	7.62	0.060
4.5	1.11	0.22	21.51	0.92	0.002	0.94	0.014	5.42	0.039
5	1.00	0.00	21.88	0.88	0.001	0.92	0.009	3.72	0.025
5.5	0.91	−0.22	22.04	0.84	0.000	0.83	0.005	2.17	0.013
6	0.83	−0.45	22.08	0.78	0.000	0.69	0.002	1.21	0.006
6.5	0.77	−0.67	22.09	0.72	0.000	0.57	0.001	0.76	0.002
7	0.71	−0.89	22.09	0.67	0.000	0.46	0.0005	0.58	0.001

It is worth noting that none of the policies reviewed here is feasible for the control problem stated in §2, which required 100% compliance with the lead-time guarantees. Although this is possible to achieve with expediting, it would require that the firm keeps track of the age of each job in the system, making the solution of the approximating MDP problem intractable.

6.1. Single Lead-Time

In order to isolate the effect of dynamic over static pricing, this subsection focuses on problems of pricing and expediting for a single product offered with a lead-time guarantee. We consider the following setup for these experiments. The demand model is that of §5, and unless otherwise stated, its parameters will be as follows: the market potential is $\Lambda_o = 10$ and $b_1 = 1$, $b_2 = 0.15$, $\gamma_0 = 2$, and $\gamma_1 = 0.4$ (cf. §5). Service times are i.i.d. exponentially distributed with rate μ. The expediting cost is $c = \$5$ per order, and expediting is used whenever the queue length reaches the threshold $\mu \cdot d - \delta$, where $\delta \geq 0$ is a "tune"-parameter, the effect of which will be studied in Table 1. In Tables 1 and 2, the offered lead time is $d = 4$, which optimizes the profit rate under static pricing. For these parameters, the capacity unconstrained revenue maximizing demand rate is $\hat{\lambda}(d) = 5$.

6.1.1. The Effect of the Expediting "Tunable"-Parameter δ. The first set of results focuses on the efficiency of the proposed expediting policy. In these experiments, the service rate is $\mu = 4$, which gives that $\hat{\rho}(d) = \hat{\lambda}(d)/\mu = 1.25$ (elaborated more in the next part), or equivalently that $\kappa = 0.45$ (cf. Equation (11)). We note that the gap between static and dynamic pricing was around 6.5%, providing an illustration of the conservative nature of static pricing policies, which, in turn, lead to higher probabilities of expediting (because price increases cannot be used to turn away orders), but lower probabilities of lead-time violation. The average tardiness is also shorter under static pricing (recall that the target lead time is $d = 4$). The effect of the tune parameter δ on the probability of violating the lead-time guarantee was as expected, and hereafter this parameter will be selected so that the probability of an order violating its lead-time guarantee is no greater than 3%; the choice of 3% is arbitrary, and serves to make our comparison more appropriate by requiring each policy to adhere to a similar lead-time violation standard. (In almost all tests the static pricing policy had a smaller parameter δ than the dynamic pricing one.)

6.1.2. The Effect of Capacity Imbalance or Load Factor ($\hat{\rho}$). The accuracy of the approximations used in §3 is higher in systems where the capacity unconstrained revenue maximizing demand rate $\hat{\lambda}(d)$ is close to the available capacity μ. This is best described by the load factor $\hat{\rho}(d) = \hat{\lambda}(d)/\mu$; although in the context of our analysis this is captured via the parameter "κ" which measures the distance $\hat{\lambda}(d) - \mu$ in multiples of $\sqrt{\hat{\lambda}(d)}$, the natural scale on which to study and control the behavior of the system. Table 2 explores the dependence of our results with respect to this parameter. We note that the dynamic pricing heuristic is most effective relative to the best static pricing policy in moderate values of κ or load factors between 0.8 and 1.2, and, in particular, their relative gap shrinks when κ or $\hat{\rho}(d)$ gets large. If $\hat{\rho}(d) \ll 1$, then the system is overcapacitated and a static pricing policy is almost optimal. If $\hat{\rho}(d) \gg 1$, the system is undercapacitated and although the dynamic pricing heuristic of §4 outperforms the best static pricing policy, it still performs poorly in comparison to the MDP solution. The latter is explained by the fact that the performance criterion of the diffusion control problem formulation was motivated by problem settings where the capacity is close to being balanced (cf. Equation (11)).

6.1.3. The Effect of the Lead Time (d). Table 3 studies the system behavior under the dynamic and static pricing heuristics derived in §3 as a function of the quoted lead time. The main observation is that the impact of dynamic pricing is more pronounced when lead times are shorter, because in such cases static prices have to be selected conservatively to avoid excessive use of expediting. As the target lead time gets large the expediting costs are reduced but so are the prices that the firm can charge to its prospective customers, reducing the overall expected profits. Moreover, the dynamic pricing capability of the

Table 3 The Effect of Different Lead-Time Guarantees

	MDP			Dynamic			Static		
d	$\mathbb{E}[\pi]$	$\mathbb{P}(exp)$	$\mathbb{P}(LT)$	$\mathbb{E}[\pi]$	$\mathbb{P}(exp)$	$\mathbb{P}(LT)$	$\mathbb{E}[\pi]$	$\mathbb{P}(exp)$	$\mathbb{P}(LT)$
2	20.33	0.022	0.024	18.63	0.115	0.030	17.20	0.152	0.002
3	20.93	0.007	0.019	19.65	0.068	0.028	18.83	0.087	0.004
4	20.83	0.003	0.017	19.83	0.046	0.027	19.24	0.060	0.004
5	20.50	0.002	0.014	19.62	0.037	0.026	19.21	0.046	0.004
6	20.05	0.001	0.012	19.30	0.029	0.024	18.97	0.037	0.005

first two policies led to an increase of the expected throughput times (to about 50% of the target lead time from 35% with static pricing), and a reduction to their variability; i.e., the overall distribution of observed throughput times was more closely clustered around—highly skewed left toward—the target lead-time d.

We also experimented with varying other parameters such as the market potential Λ_o and the expediting cost c. The former has a similar effect to decreasing the capacity μ (cf. Table 2), whereas the latter had the expected effect that as c increases, prices increase so as to lower the probability of expediting, and in such cases the benefits from dynamic pricing are more important.

6.2. Effect of Lead-Time Flexibility: Single Good Offered at Two Lead Times

We adopt the same model parameters as for the experiments reported above, setting the service rate at $\mu = 4$, together with the specification that the delay sensitivity parameter χ used in selecting a product in (37) is uniformly distributed in an interval $[0, \chi_m]$, and $\chi_m = 2$. Since both products correspond to the same good, we will assume that $c_i = 5$ for $i = 1, 2$.

Our first set of results looks at the impact of lead-time flexibility on a baseline example that was already analyzed in the previous subsection, for which $\mu = 4$ and a single lead-time option was offered at $d = 4$ (that corresponds to the first row in Table 2). The results in this table study the performance of dynamic and static pricing policies for various pairs of lead-times (d_1, d_2). First, we note that lead-time flexibility leads to a 5% to 8% performance improvement for the MDP and dynamic pricing policies when compared to the performance under the same policies with the single lead-time option. The performance gain due to lead-time flexibility was larger (8% to 10%) under the static pricing policy. Adding lead-time flexibility leads to a reduction of the performance gaps between the dynamic and static pricing heuristics and the MDP solution; the results of Table 4 correspond to the first row in Table 2, where the two gaps were 4.80% and 7.62%, respectively.

We complement this table by reporting average results from a larger set of test problems where we varied some of the demand model parameters as follows: $\Lambda_o \in \{8, 10, 12\}$, $b_1 \in \{0.8, 1, 1.2\}$, and $b_2 \in \{0.1, 0.15, 0.2\}$. For all possible parameter combinations, we first considered the problem of offering one product option and searched for the optimal lead-time d^* under the static pricing policy. We then tested the performance of the dynamic and static pricing policies for the case where the firm offered two lead-time options defined as $d_1 = (0.8)d^*$ and $d_2 = (1.2)d^*$. Once again, we report the percentage of profit gains over the single lead-time system under static pricing. We observed the following results:

(i) Dynamic pricing: average profit gain was 13.97% with a standard deviation of 2.34%.

(ii) Static pricing: average profit gain was 10.69% with a standard deviation of 2.99%.

(iii) Dynamic versus static pricing: average gap 3.28% with a standard deviation of 2.03%. Actual differences ranged in [1.06%, 10.30%].

Table 4 Performance Measures for Different Lead-Time Combinations

	MDP		Dynamic		Static	
d_1, d_2	$\mathbb{E}[\pi]$	$\Delta(\pi_{\text{MDP}})$ (%)	Gap (%)	$\Delta(\pi_S)$ (%)	Gap (%)	$\Delta(\pi_S)$ (%)
3, 4	22.04	5.49	3.37	9.65	5.15	7.95
3.5, 4	22.51	7.47	4.66	10.34	6.18	8.88
3.5, 4.5	22.42	7.09	2.46	12.00	5.92	8.77
3.5, 5	22.24	6.35	3.22	10.61	4.91	9.02
4, 4.5	22.40	7.00	1.58	12.71	4.72	9.84
4, 5	22.64	8.00	3.71	11.74	5.35	10.21

Note. $\Delta(\pi_{\text{MDP}})$ is the percentage of gain over the single lead-time MDP solution; Gap percent is the performance gap relative to the MDP solution with two lead times; and $\Delta(\pi_S)$ is the percentage of gain relative to the single lead-time static pricing policy.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Acknowledgments

The authors are grateful to Philipp Afèche, Baris Ata, and Omar Besbes, the associate editor, and two referees for their helpful comments. The second author's research was partially supported through a grant from the Columbia Center for Excellence in E-Business.

References

Afèche, P. 2004. Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delaying tactics. Working paper, Kellogg School of Management, Northwestern University, Evanston, IL.

Ata, B. 2003. Dynamic control for stochastic networks. Ph.D. thesis, Graduate School of Business, Stanford University, Stanford, CA.

Ata, B. 2006. Dynamic control of a mutliclass queue with thin arrival streams. *Oper. Res.* **54**(5) 876–892.

Ata, B., J. M. Harrison, L. A. Shepp. 2005. Drift rate control of a Brownian processing system. *Ann. Appl. Probab.* **15**(2) 1145–1160.

Baker, K. 1984. Sequencing rules and due-date assignmenets in a job shop. *Management Sci.* **30**(9) 1093–1104.

Bradley, J. L. 2004. A Brownian approximation of a production-inventory system with a manufacturer that subcontracts. *Oper. Res.* **52**(5) 765–784.

Bradley, J. R. 2005. Optimal control of a dual service rate $M/M/1$ production-inventory model. *Eur. J. Oper. Res.* **161**(3) 812–837.

Brémaud, P. 1980. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, New York.

Bucklin, R. E., S. Gupta. 1992. Brand choice, purchase incidence, and segmentation: An integrated modeling approach. *J. Marketing Res.* **29**(2) 201–215.

Charnsirisakskul, K., P. M. Griffin, P. Keskinocak. 2006. Pricing and scheduling decisions with leadtime flexibility. *Eur. J. Oper. Res.* **171**(1) 153–169.

Duenyas, I. 1995. Single facility due date setting with multiple customer classes. *Management Sci.* **41**(4) 608–619.

Duenyas, I., W. J. Hopp. 1995. Quoting customer lead times. *Management Sci.* **41**(1) 43–57.

Edmondson, G. 2003. Customization—BMW. *BusinessWeek* (November 24) 94.

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

Gallego, G., G. van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Sci.* **40**(8) 999–1020.

Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems*. John Wiley & Sons, New York.

Harrison, J. M. 1988. Brownian models of queueing networks with heterogeneous customer populations. 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, New York, 147–186.

Harrison, J. M. 2003. A broader view of Brownian networks. *Ann. Appl. Probab.* **13** 1119–1150.

Harrison, J. M., J. A. Van Mieghem. 1996. Dynamic control of Brownian networks: State space collapse and equivalent workload formulations. *Ann. Appl. Probab.* **7** 747–771.

Keskinocak, P., S. Tayur. 2003. Due date management policies. D. Simchi-Levi, S. D. Wu, Z. M. Shen, eds. *Supply Chain Analysis in the e-Business Era*. Kluwer Acedemic Publishers, Norwell, MA, 485–553.

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

Lariviere, M., J. A. Van Mieghem. 2004. Strategically seeking service: How competition can generate Poisson arrivals. *Manufacturing Service Oper. Management* **6**(1) 23–40.

Maglaras, C. 2006. Revenue management for a multiclass single-server queue via a fluid model analysis. *Oper. Res.* **54**(5) 914–932.

Maglaras, C., J. A. Van Mieghem. 2005. Queueing systems with leadtime constraints: A fluid-model approach for admission and sequencing control. *Eur. J. Oper. Res.* **167**(1) 179–207.

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.

Maglaras, C., A. Zeevi. 2004. Design and performance analysis of differentiated services with customer choice. Working paper, Columbia University, New York.

Mandelbaum, A., G. Pats. 1995. State-dependent queues: Approximations and applications. F. Kelly, R. Williams, eds. *Stochastic Networks*, Vol. 71, IMA Volumes in Mathematics and its Applications. Springer-Verlag, New York, 239–282.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Comm. ACM* **28**(3) 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** 15–24.

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

Plambeck, E. L., A. R. Ward. 2008. A separation principle for assemble-to-order systems with expediting. *Oper. Res.* Forthcoming.

Plambeck, E., S. Kumar, J. M. Harrison. 2001. Leadtime constraints in stochastic processing networks under heavy traffic conditions. *Queueing Systems* **39** 23–54.

Reiman, M. I. 1984. Open queueing networks in heavy traffic. *Math. Oper. Res.* **9**(3) 441–458.

Reiman, M. I. 1988. A multiclass feedback queue in heavy traffic. *Adv. Appl. Probab.* **20**(1) 179–207.

Rubino, M., B. Ata. 2008. Dynamic control of a make-to-order, parallel-server system with cancellations. *Oper. Res.* Forthcoming.

Talluri, K., G. J. van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Springer, New York.

Wein, L. M. 1991. Due-date setting and priority sequencing in a multiclass $M/G/1$ queue. *Management Sci.* **37**(7) 834–850.