Near-Optimal Dynamic Lead-Time Quotation and Scheduling Under Convex-Concave Customer Delay Costs

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

Tava Lennon Olsen
Washington University in St. Louis, St. Louis, Missouri 63130,
olsen@olin.wustl.edu

We consider a make-to-order system where customers are dynamically quoted lead times (and prices). Customers are homogenous but have general (nonlinear) disutility for delay. Because the firm is a monopolist, the pricing problem is trivial and the dynamic problem reduces to one of lead-time quotation and order sequencing. We also consider the (static) problem of up-front capacity installation. We use a large-capacity asymptotic regime to make the problem tractable. We provide recommended policies for convex, concave, and convex-concave lead-time cost functions and prove that these policies are asymptotically optimal. The policies are both highly intuitive and readily implementable. Moreover, they provide delay guarantees for all served customers. They are tested numerically; we find that significant benefits can accrue by using the prescribed dynamic policies instead of first-come-first-served type policies.

Subject classifications: production/scheduling: dynamic lead-time quotation; queues: limit theorems.
Area of review: Manufacturing, Service, and Supply Chain Operations.
History: Received January 2007; revisions received September 2007, January 2008; accepted March 2008. Published online in Articles in Advance March 11, 2009.

1. Introduction

“Time is money,” as Benjamin Franklin famously said. Here we assume not only that this is the case, but that costs are not necessarily linear in time lengths. In particular, we consider lead-time quotation in a make-to-order system where customers have nonlinear disutilities on the delays they are quoted. The system manager tells each customer up front what maximum delay or lead time she guarantees they will receive; if the benefits of service do not exceed a customer’s disutility for waiting then he will not place an order. The system manager must therefore dynamically determine what lead time she should quote (at what price) to each arriving customer and how to schedule customers present in the system. She must also determine what up-front capacity to install.

We consider a monopolist who may extract the entire customer’s utility for decreased lead time. Thus, the price for a given lead time is the price for the maximum acceptable lead time plus the customer’s delay cost savings for the shorter lead times. The firm seeks to maximize its revenue minus the cost of capacity. In this case the pricing problem has been effectively eliminated and the dynamic problem reduces to one of lead-time or due-date quotation and order sequencing to meet these lead times. Although such a monopolist is not the most general setting that may be imagined, it provides the cleanest and simplest modeling environment to examine the effect that the shape of the delay cost function has on lead-time quotation and order sequencing, which is our primary contribution.

As described above, we seek to explicitly study the effect that the shape of the disutility or delay cost function has on the firm’s policies. Therefore, to isolate this effect, we assume homogeneous customers; that is, all customers have the same (deterministic) delay cost function. Under this assumption, we find that it is asymptotically optimal to specify a threshold beyond which customers are not admitted (or, equivalently, quoted a lead time that is unacceptably long resulting in their choosing not to join the system). Furthermore, both for tractability and to isolate the effect of the shape of the cost function, we also do not consider competition.

We consider three possible shapes for lead-time cost curves, namely, convex, concave, and convex-concave. A convex curve is an appropriate model for situations where customers have a general idea on desirable lead time and longer lead times than this are increasingly unattractive. Lead times shorter than a customer’s a priori acceptable lead time are attractive but not unduly so. A concave cost curve represents the opposite; namely, there are decreasing costs to increasing lead times (i.e., customers have strong
preferences for lead times that are as short as possible and are increasingly indifferent to longer lead times). A convex-concave, or “S-shaped,” cost curve is a hybrid where customers may have a particular deadline in mind (e.g., a spouse’s birthday) but once that deadline has passed there is increasingly little difference in the added lead time. Obviously, it encompasses convex and concave cases as subcases.

One innovation of our work is a shift-based approach to lead-time quotation. Lead times are quoted to the nearest shift (rather than in arbitrarily small units), and once a lead time is quoted it must be met. We do not seek to model the trade-off between short lead times and higher tardiness costs. Instead, we seek to model the trade-off between committing future capacity now and reserving it for potential later higher-revenue customers. The former trade-off may then be modeled in setting quotas for a shift. However, it is the latter that we feel is more important when modeling the effect that the shape of the lead-time cost curve has on the optimal policy. Although our recommended policies are readily implementable, we feel that their most significant contribution is the intuition that they provide into the effect of the shape of the cost curve.

When costs are convex we find that first-come-first-served (FCFS) sequencing is optimal (this echoes existing results in the literature—see §2 for an overview). However, when costs are not convex then FCFS is well known not to be optimal and may, in fact, be far from optimal (see §7 for numerical examples). We believe our work to be the first to propose near-optimal, easily implementable lead-time quotation and sequencing policies for nonconvex lead-time cost curves.

For nonconvex lead-time costs, the convex hull of the cost function serves as a lower bound to the system costs. The key idea behind the policies we propose for concave and convex-concave costs is to asymptotically approach the cost incurred by the convex hull of the cost function. In these two cases, the proposed lead-time quotation and scheduling policy is specified by a high-priority service capacity and a low-priority service capacity. The segmentation is arbitrary, but what results in the concave case is most customers receiving a lead time of one shift while a long thin tail of low-priority customers are kept as a hedge against uncertainty in the arrival process. Figure 1 depicts a typical workload distribution under the proposed policy for concave costs.

Figure 1 lends further illumination to the trade-off discussed above between committing future capacity now and reserving it for potential later higher-revenue customers. Although reserving capacity penalizes those customers quoted a long lead time, it also frees up capacity for newly arriving customers who may then be quoted a lead time of one. In the case of convex-concave cost curves, the segmentation is less extreme but the idea is similar. The majority of customers are given relatively short lead times, but when system workload reaches a sufficient level of congestion a small portion of customers are given very long lead times and kept as a hedge against demand uncertainty while allowing most customers to receive moderate lead times.

The proposed policies are intuitively appealing. They are also very easy to implement. The policies would be even more appealing if one could perform customer segmentation so that the long-lead-time customers are those that are truly the more patient customers. Modeling heterogeneous customers is left as the subject of future research.

Our model will assume that lead time depends on congestion; therefore, the lead time associated with shipping is assumed to be on top of (and independent of) the quoted lead time. In other words, customers are quoted a lead time for the time until the product leaves the factory (or the service is complete), and delivery time is handled independently. Because production is make-to-order, lead time is the flow time for the order. In this case, the “product” may be highly customized (e.g., furniture production), but the service time distribution must be the same across customers.

We consider systems with a high volume of arrivals. Because of the statistical economies of scale phenomenon, such systems become more and more efficient as they grow, and the congestion concerns become less important for larger systems. As a result, one can operate such high-volume systems near full utilization and yet achieve excellent performance. Indeed, our analysis also validates this intuition; cf. Proposition 1.

The remainder of this paper is organized as follows. Section 2 contains a brief literature review. In §3, we outline our model and present some preliminary results. Section 4 presents the asymptotic optimization problem and develops an asymptotic upper bound for system performance. Section 5 solves the asymptotic control problem for the upper bound on performance. This upper bound is then used to prove that the proposed control policies presented in §6 are asymptotically optimal; these policies are tested numerically in §7. The paper is concluded in §8. The proofs are relegated to an online technical appendix that can be found at http://or.journal.informs.org/.

**Figure 1.** An illustration of the workload distribution between the two classes under the proposed policy in the concave case.
2. Literature Review

Our work is probably most related to the due-date quotation literature, which, of course, has a long history. Hopp and Sturgis (2000) contains a relatively recent review of this literature. In their paper they seek to quote the shortest possible lead time consistent with a given service constraint. Their model and most others assume demand to be independent of lead time. For example, in both Spearman and Zang (1999) and Wein (1991) the supplier seeks to minimize the average due-date lead time subject to a constraint on job tardiness. Also see Keskinocak and Tayur (2004) for an overview of due-date management policies and Cheng and Gupta (1989) for a survey of scheduling research involving due-date determination decisions.

In our model, the customer either accepts the quoted lead time or leaves (i.e., is turned away by being quoted too high a lead time). A related approach is taken in Duenyas and Hopp (1995) and Duenyas (1995). In their continuous-time models, if a customer is quoted a lead time of \( a \), then he will accept with probability \( \pi(a) \). There is a net revenue per customer and a (linear) penalty \( cx \) if the customer is delayed for \( x \) units of time.

Due-date quotation when demand decreases with increased lead time has been addressed by a number of other authors. In particular, Armony and Maglaras (2004) consider a model of call centers where customers are quoted a lead time and may either leave or place a request for a call back within a prespecified (different) lead time. Chatterjee et al. (2002) consider the marketing aspects of lead-time quotation; they also provide a nice summary of the literature (see their Table 1). Duran et al. (2006) have a model where poor lead-time performance can lead to decreased future demand. Rao et al. (2005) consider a model where mean demand in a period decreases with lead time whereas the stochastic component of demand is independent of lead time. Quoted lead time is always met by the customer either accepts the quoted lead time or leaves (i.e., is turned away by being quoted too high a lead time). A related approach is taken in Duenyas and Hopp (1995) and Duenyas (1995). In their continuous-time models, if a customer is quoted a lead time of \( a \), then he will accept with probability \( \pi(a) \). There is a net revenue per customer and a (linear) penalty \( cx \) if the customer is delayed for \( x \) units of time.

Table 1. Impact of system scale and the excess capacity (drift) on the percentage improvement of the proposed policy over the benchmark policy in the concave case.

<table>
<thead>
<tr>
<th>( \nu )</th>
<th>( n = 100 )</th>
<th>( n = 400 )</th>
<th>( n = 1,000 )</th>
<th>( n = 6,400 )</th>
<th>( n = 10,000 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 3 )</td>
<td>14.21 ± 0.75</td>
<td>21.49 ± 1.40</td>
<td>25.61 ± 1.99</td>
<td>32.40 ± 4.05</td>
<td>32.86 ± 4.70</td>
</tr>
<tr>
<td>( 5 )</td>
<td>19.82 ± 0.51</td>
<td>25.89 ± 0.97</td>
<td>29.05 ± 1.62</td>
<td>34.40 ± 3.08</td>
<td>34.80 ± 4.06</td>
</tr>
<tr>
<td>( 7 )</td>
<td>25.79 ± 0.33</td>
<td>30.98 ± 0.74</td>
<td>33.92 ± 1.19</td>
<td>38.62 ± 2.49</td>
<td>38.76 ± 3.21</td>
</tr>
</tbody>
</table>

We do not consider competition. There are a variety of game-theoretic price and service models (see, e.g., Afèche and Mendelson 2004, Bashyam 2000, Hassin and Haviv 2003, Lederer and Li 1997, Li and Lee 1994, So 2000). We also do not consider any informational effects with respect to lead-time quotation; customers are quoted guaranteed lead times, which they either accept or reject. In a recent series of papers Guo and Zipkin (2007a, b, c) model balking behavior that depends on lead-time quotes (see also references therein).

The marketing literature contains considerable work on measuring waiting costs and utilities, but we are aware of no modeling work. Some relevant recent references are as follows. Antonides et al. (2002) perform an experiment to evaluate people’s perceptions of waiting time. Frederick et al. (2002) give a review of intertemporal discounting. Other studies in the literature include anomalies in people’s preferences and differences between savings and costs (see, e.g., Kahneman et al. 1990) or between the status quo and other options (see, e.g., Samuelson and Zeckhauser 1988 or Kahneman et al. 1991). Here we simply assume that there is a known and fixed cost for waiting.

We are aware of no literature that explicitly addresses the most applicable shape for the lead-time cost curve. This could be because such a question is relevant only if one chooses to use the resultant curve shape explicitly in a model. Furthermore, it is likely that the shape of the curve is highly application specific. However, we would note that S-shaped curves are used elsewhere in the marketing literature to model consumer responses; for example, S-shaped curves are used as response functions to advertising (e.g., Sasieni 1971).

One primary contribution of our work is the (asymptotically optimal) sequencing of jobs with nonlinear delay cost functions. There is relatively little work in the scheduling and sequencing literature that deals with nonlinear and, particularly, nonconvex costs. More common metrics used are mean throughput time, mean cost per unit time, or in the due-date scheduling literature (e.g., Wein 1991) a linear function of due dates and arrival times. However, a number of authors have considered convex cost functions (e.g., Van Mieghem 2000). It is well known that with a single class FCFS is optimal under nondecreasing convex costs (e.g., Wolff 1989, ex. 5.41), a result we also establish for the specific model considered here.

There has also been work on minimizing the variance of delay (a convex but not increasing objective). In the deterministic scheduling literature this gives rise to a “V-shaped” schedule where longer jobs are scheduled both at the beginning and at the end of the schedule (see Eilon and Chowdhury 1977). Mittenthal et al. (1995) provide a more
general cost function that is the sum of three functions of job completion times: (i) the sum of the squares, (ii) the square of the mean, and (iii) the mean. A variety of well known scheduling rules (e.g., V-shaped, shortest-processing-time, etc.) are special cases of this objective. V-shaped schedules also appear with earliness and tardiness penalties (see Baker and Scudder 1990 for a review of this literature). Research on minimizing delay variance in dynamic (nondeterministic) systems is surveyed in Ayhan and Olsen (2000).

As described in the introduction, we will consider both concave and convex-concave delay cost functions. The only scheduling work we can find that explicitly considers concave cost functions is Hain and Mitra (2004). They present a game-theoretic framework where all jobs are present at time zero and the (independent) agents must be induced to truthfully reveal their job processing times. We can find no scheduling work that explicitly considers convex-concave cost functions.

There have been a few papers that explicitly consider both pricing and lead-time decisions, the most related of which are as follows. Çelik and Maglaras (2008) provide an approximating diffusion control problem, like we do, for an environment where lead times are precommitted and the arrival rate depends on the prices quoted. Keskinocak et al. (2001) model revenue-sensitive lead time. They have a model with deterministic service and bound performance of the system using competitive analysis. In Van Mieghem (2000), customers sign up to differing service grades with the right to send a stream of service requests over times. The grades are differentiated by expected lead time. The price schedule is static; however, the actual price paid may be a function of the observed processing time and the chosen grade. Plambeck (2004) considers static prices for each of two classes of customer where lead times are quoted dynamically. Arrival rates decrease with price and lead time, and the object is to maximize profit subject to meeting the promised lead time. Heavy-traffic analysis is used to analyze the system and arrive at an asymptotically optimal policy.

Finally, other related work, particularly methodologically, includes Gallien et al. (2005), Maglaras and Zeevi (2005), and Kumar and Randhawa (2008). Gallien et al. (2005) consider a model of dynamic admission control for a make-to-order system where customers are differentiated by price, quantity, and lead time. Lead times are assumed fixed. Maglaras and Zeevi (2005) consider a model where some users are guaranteed a processing rate while others are best-effort. Users are sensitive to both price and lead time. The paper presents pricing and admission rules that are “fluid-optimal.” Kumar and Randhawa (2008) consider revenue optimization with nonlinear, convex delay costs.

3. The Model

As mentioned in the introduction, we consider a single class of (homogeneous) customers arriving to a make-to-order monopolist firm, which we model as a single-server queue. The firm chooses its capacity at the beginning of the planning horizon, and a system manager quotes lead times (and prices) to arriving customers dynamically over time. In addition to capacity and lead-time quotation decisions, the system manager also chooses the sequence in which the customer orders are processed so that the quoted lead times are respected.

Each customer has a delay cost of $c(\tau)$ associated with receiving a lead-time quotation of $\tau$ and receives benefit $R$ from service. Shorter lead times are more attractive to customers; i.e., the delay cost function $c(\tau)$ is increasing in the quoted lead time $\tau$; a customer will join the system only if $c(\tau) \leq R$ when he is quoted a lead time of $\tau$. In other words, the system manager can induce a customer not to join the system by quoting a large lead time even though it is the customer himself making the decision of whether or not to join the system.

Adopting the terminology that is standard in queueing theory, we refer to inducing a customer not to join the system as turning away a customer and view the system manager as making admission control decisions. By the same convention we will refer to customer orders to be processed as the jobs in the system. The number of customers turned away until time $t$ is denoted by $U(t)$. Moreover, the system manager charges price $R - c(\tau)$ to a customer joining the system whenever she quotes him a lead time of $\tau$, thereby extracting his entire surplus. Payments are collected upon service completion.

The system is initially empty, and customers arrive to the system at rate $\lambda$ according to a renewal process $\{A(t), t \geq 0\}$, where the coefficient of variation of the interarrival times is $\alpha > 0$. To be more specific, letting $\{v_i; i \geq 1\}$ be a sequence of independent and identically distributed random variables with mean one and variance $\sigma^2$, the cumulative number $A(t)$ of arrivals up to time $t$ is given by

$$A(t) = \max \{l \geq 0; \sum_{i=1}^{l} v_i \leq \lambda t \}, \quad t \geq 0.$$  

We assume that the interarrival times have exponential moments in a neighborhood of the origin.

To guarantee that the quoted lead time will always be met, we restrict attention to deterministic production. Deterministic production is a good model of systems with a fixed capacity (e.g., a movie theater) or where production is according to a fixed quota. There are many environments where overtime is run at the end of the day (if needed) and therefore the number of items to be produced within a day may be assumed to be constant. For example, the traditional Toyota production system is based on two shifts with overtime as a stopgap between shifts (see, e.g., Schonberger 1982).

In our model, the firm chooses its capacity at the beginning of the planning horizon by choosing its service rate $\mu$. The production decisions are made at discrete points in time, say, at times $0, \kappa, 2\kappa, \ldots$; and customers are quoted
lead times that are integer multiples of \( \kappa \). One can view our policy as a discrete review policy, where review periods correspond to production shifts and \( \kappa \) denotes the length of a production shift. At the beginning of every review period, the system manager monitors the system status and chooses how many customer orders to admit during the current period (which will be processed starting only in the next period). As the system manager admits a customer, she quotes him a lead time (and a corresponding price).

The production plan for each shift concerns processing of only the jobs that are present at the beginning of the shift. In particular, the production plan is frozen until the beginning of the next shift, and the jobs arriving during the shift are not processed until the next shift begins even if there is spare capacity in the current shift, which does not cause any significant performance degradation for large systems, as will be shown below, provided that the system manager reviews the system status sufficiently often. Furthermore, we restrict attention to policies that are nonidling because the admission control policy by \( \varphi = (\mu, U(-), \{ \tau_k \; k \geq 1 \}) \). For admissibility, we require that the system manager’s policy be nonanticipating in the usual sense. Although we will propose a policy that makes production scheduling decisions within the discrete-review framework in each of the three cases we will consider, the class of admissible policies we allow includes policies in which the system manager can make all decisions continuously over time, allowing in particular to work on the jobs that arrive in the current period. Given an admissible policy, the cumulative revenue up to time \( t \) is given by

\[
\sum_{i=1}^{A(t)-U(t)} (R - c(\tau_i)) = RA(t) - RU(t) - \sum_{i=1}^{A(t)-U(t)} c(\tau_i).
\]

The capacity cost up to time \( t \) is \( \nu \mu t \), where \( \nu < R \) is the cost of unit capacity per time unit. Therefore, cumulative profits under policy \( \varphi \), denoted by \( \Pi_\varphi(t) \), are given by

\[
\Pi_\varphi(t) = RA(t) - RU(t) - \sum_{i=1}^{A(t)-U(t)} c(\tau_i) - \nu \mu t.
\]

The objective is to maximize long-run average profit. If the cost function \( c(\cdot) \) is convex then this problem is highly tractable, at least in terms of lead-time quotation, and a sample path argument yields the result that customers should be quoted lead times on a FCFS basis; this is given in Theorem 1 where its proof (and all following proofs) is (are) given in the online appendix. The optimality of FCFS is not at all surprising given previously established results in the literature (e.g., Van Mieghem 2000, Wolff 1989). For concave and convex-concave cost functions, the problem does not appear to be analytically tractable, and we use a bounding approach combined with large-capacity asymptotics, which will be detailed in the following sections.

**Theorem 1.** If customers have homogeneous convex cost functions for lead time, then, given a customer is accepted, he will be quoted the shortest available (feasible) lead time.

Observe that, although Theorem 1 specifies the optimal lead-time quotation policy for the admitted customers, it says nothing about the admission control policy. Neither does it say anything about the optimal capacity decision.

In §§4 and 5, we develop an asymptotic upper bound on the system performance for the following three cases: (i) \( c(\cdot) \) is convex; (ii) \( c(\cdot) \) is concave; (iii) \( c(\cdot) \) is convex-concave, which corresponds to \( c(\cdot) \) being convex on an interval \([0, \bar{x}]\) and concave on \([\bar{x}, \infty)\). To enhance the reader’s intuition, we next illustrate the main idea of our approach to develop the performance bound. Our analysis provides a unified treatment of the three possible shapes of the delay cost function \( c(\cdot) \) in terms of the convex hull \( h(\cdot) \) of \( c(\cdot) \). The convex hull of \( c(\cdot) \) is given by the maximal convex function \( h \leq c \); cf. Rockafellar (1970). Clearly, replacing the delay cost \( c(\cdot) \) by its convex hull \( h(\cdot) \) provides a lower bound on the system delay costs for any given admissible policy. Moreover, because \( h(\cdot) \) is convex, it is optimal in this new system to serve the admitted customers on a FCFS basis and quote lead times accordingly. This simple observation is the essence of our approach.

To facilitate our analysis, define the workload \( W(t) \) in the system at time \( t \) as follows:

\[
W(t) = A(t) - \mu t + L(t) - U(t), \quad t \geq 0,
\]

where \( L(t) \) is the cumulative unused capacity up to time \( t \). For any fixed admission control policy \( U \), the evolution of the workload process is the same for all work conserving policies. Because the admitted customers are served on an FCFS basis in this new system with the delay cost function \( h(\cdot) \), the customer arriving at time \( t \) is quoted a lead time of \( W(t)/\mu \); cf. Theorem 1. Thus, an upper bound on \( \Pi_\varphi(t) \) is given by

\[
\xi(t) = RA(t) - RU(t) - \int_0^t h\left( \frac{W(s)}{\mu} \right) d(A(s) - U(s)) - \nu \mu t.
\]
Moreover, the server keeps working as long as there are jobs in the system. In other words, the idleness process $L$ increases only when the workload in the system is zero. That is, the following complementarity condition holds for $t \geq 0$:

$$
\int_0^t 1_{[w(s) > 0]} \, dL(s) = 0.
$$

One can further maximize the upper bound (1) by choosing the optimal capacity and the admission control policy $U$ to get an upper bound on the optimal performance of the original system. Unfortunately, that optimization problem is not tractable analytically. Therefore, we consider a sequence of systems in the large-capacity asymptotics and arrive at a far more tractable formulation in the next section.

4. Large-Capacity Asymptotics

In this section, we introduce a sequence of systems indexed by $n = 1, 2, \ldots$. A superscript $n$ will be attached to the quantities corresponding to the $n$th system in this sequence. To be specific, the asymptotic regime we are interested in is the one where the arrival rate $\lambda^n$ grows with $n$. Namely, we assume that

$$
\lambda^n = n \quad \text{for} \ n \geq 1.
$$

In this asymptotic regime, our goal is to construct a sequence of policies $\{\nu^n; n \geq 1\}$ (one for each system) that is asymptotically optimal in a sense to be made precise shortly.

To facilitate our definition of asymptotic optimality, we first consider an idealized static planning problem. Suppose that the system manager ignores the stochastic variability in customer arrivals and queueing delays caused by this variability. To maximize the long-run average profit rate, she will quote zero lead time, charge price $R$ to every customer, and set capacity according to the following static planning problem:

$$
\Pi^n = \max_{\mu \geq \lambda^n} R\lambda^n - n\mu. \tag{2}
$$

Clearly, one must set $\mu = \lambda^n$ so that $\Pi^n = n(R - \nu)$, which is an upper bound on the long-run average profit rate in the $n$th system. We would like to construct a sequence of policies $\{\nu^n; n \geq 1\}$ that satisfies

$$
\lim_{\mu \to \lambda^n} \lim_{n \to \infty} \frac{1}{nT} E[\Pi^n_\varphi(T) - T\Pi^n] = 0.
$$

That is, we want to show that

$$
\lim_{\mu \to \lambda^n} \lim_{n \to \infty} \frac{1}{nT} E[\Pi^n_\varphi(T)] = (R - \nu).
$$

Better yet, we would like to maximize the rate of convergence of the average profit to its upper bound by constructing a sequence of policies $\{\nu^n; n \geq 1\}$ that satisfies

$$
\lim_{\mu \to \lambda^n} \lim_{n \to \infty} \frac{1}{\sqrt{n}} \left( \frac{1}{nT} E[\Pi^n_\varphi(T)] - (R - \nu) \right) \geq \lim_{\mu \to \lambda^n} \lim_{n \to \infty} \frac{1}{\sqrt{n}} \left( \frac{1}{nT} E[\Pi^n_\varphi(T)] - (R - \nu) \right)
$$

for any other sequence of admissible policies $\{\varphi_n; n \geq 1\}$.

Defining the diffusion-scaled profits $\bar{\Pi}^n_\varphi(t)$ up to time $t$ under policy $\varphi$ in the $n$th system as

$$
\bar{\Pi}^n_\varphi(t) = \frac{1}{\sqrt{n}} \left[ \Pi^n_\varphi(t) - n(R - \nu) t, \quad t \geq 0,
$$

the preceding condition can be written in terms of the diffusion-scaled profit rate as follows:

$$
\lim_{\mu \to \lambda^n} \lim_{n \to \infty} \frac{1}{T} E[\bar{\Pi}^n_\varphi(T)] \geq \lim_{\mu \to \lambda^n} \lim_{n \to \infty} \frac{1}{T} E[\bar{\Pi}^n_\varphi(T)]. \tag{3}
$$

Therefore, we adopt the following definition of asymptotic optimality.

**Definition 1.** A sequence of policies $\{\nu^n; n \geq 1\}$ is called asymptotically optimal if for any other sequence of admissible policies $\{\varphi_n; n \geq 1\}$ it satisfies

$$
\lim_{\mu \to \lambda^n} \lim_{n \to \infty} \frac{1}{T} E[\bar{\Pi}^n_\varphi(T)] \geq \lim_{\mu \to \lambda^n} \lim_{n \to \infty} \frac{1}{T} E[\bar{\Pi}^n_\varphi(T)].
$$

The following proposition characterizes the economically optimal order of magnitude for the system capacity.

**Proposition 1.** If $\lim_{n \to \infty} |\lambda^n - \mu^n|/\sqrt{n} = \infty$, then, under any sequence $\{\nu^n; n \geq 1\}$ of admissible policies, we have that

$$
\lim_{n \to \infty} \frac{1}{T} E[\bar{\Pi}^n_\varphi(T)] = -\infty.
$$

Motivated by Proposition 1, we will restrict attention to policies that satisfy

$$
\lim_{n \to \infty} (\lambda^n - \mu^n)/\sqrt{n} = \theta, \quad \text{where} -\infty < \theta < \infty.
$$

which reduces the problem of capacity sizing to choosing the second-order adjustment $\theta\sqrt{n}$ to the nominal capacity suggested by the static planning Problem (2).

Equation (4) is the familiar heavy-traffic assumption (cf. Harrison 1988), but here we find that heavy traffic emerges as a consequence of optimal capacity decisions as in Plambeck (2004) and Maglaras and Zeevi (2003, 2005). Moreover, under condition (4), the optimal capacity choice gives rise to a large balanced-flow system. For such large balanced-flow systems, the workload in the system is expected to be in the order of $\sqrt{n}$, whereas we expect...
delays to be in the order of $1/\sqrt{n}$. Thus, we scale the cost of delay as follows:

$$c_n(t) = \frac{c(\sqrt{n})}{\sqrt{n}} \quad \text{for } n \geq 1.$$  

Next, we pass to the limit formally as $n \to \infty$ and derive a singular control problem, whose solution (derived in §5) provides an asymptotic upper bound on the performance of any sequence of admissible policies (cf. Theorem 3). It is important to point out that we will prove the asymptotic optimality results rigorously; and our proofs do not hinge on this formal derivation, which is provided to motivate our solution approach because it avoids various technical intricacies, making the underlying ideas more transparent.

As illustrated in §3, we will use the convex hull of the delay cost function to advance an asymptotic upper bound. It is easy to check that the convex hull $h_n(\cdot)$ of $c_n(\cdot)$ is given by the following:

$$h_n(t) = \frac{h(\sqrt{n})}{\sqrt{n}} \quad \text{for } n \geq 1.$$  

It follows from the functional strong approximations (cf. Chen and Yao 2001) that

$$A^n(t) = nt + \sqrt{n}B(t) + o(\sqrt{n}), \quad (5)$$

where $B$ is a $(0, \sigma)$ Brownian motion and $o(\sqrt{n})/\sqrt{n} \to 0$ as $n \to \infty$. Recall that the evolution of the workload in the $n$th system is governed by the following equation:

$$W^n(t) = A^n(t) - \mu^n t + L^n(t) - U^n(t), \quad t \geq 0.$$  

Combining this with (5) gives

$$W^n(t) = \sqrt{n}B(t) + \sqrt{n}\theta t + L^n(t) - U^n(t) + o(\sqrt{n}). \quad (6)$$

Similarly, the upper bound on the cumulative profit for the $n$th system is given by

$$\xi^n(t) = R(nt + \sqrt{n}B(t) + o(\sqrt{n})) - \nu \mu^n t$$

$$- \left[ RU^n(t) + \int_0^t h_n\left( \frac{W^n(s)}{\mu^n} \right) d(A^n(s) - U^n(s)) \right]. \quad (7)$$

To pass to a formal limit in (6)–(7), we introduce the following scaled quantities:

$$\hat{U}^n(t) = \frac{U^n(t)}{\sqrt{n}}, \quad \hat{L}^n(t) = \frac{L^n(t)}{\sqrt{n}}, \quad \hat{W}^n(t) = \frac{W^n(t)}{\sqrt{n}},$$

and

$$\hat{\xi}^n(t) = \frac{\xi^n(t) - (R - \nu)nt}{\sqrt{n}}.$$  

Then, assuming $\hat{U}^n \to \hat{U}$ and $\hat{L}^n \to \hat{L}$ as $n \to \infty$, scaling Equations (6)–(7) by $\sqrt{n}$, and passing to the limit formally, we conclude that

$$\hat{W}^n \to \hat{W} \quad \text{and} \quad \hat{\xi}^n \to \hat{\xi} \quad \text{as } n \to \infty,$$

where the workload process $\hat{W}$ satisfies

$$\hat{W}(t) = B(t) + \theta t + \hat{L}(t) - \hat{U}(t), \quad t \geq 0.$$  

To derive the asymptotic upper bound process $\hat{\xi}$, note that

$$\hat{\xi}^n(t) = RB(t) + R\frac{o(\sqrt{n})}{\sqrt{n}} + \nu\theta t - R\hat{U}^n(t)$$

$$- \int_0^t h\left( \frac{\hat{W}^n(s)}{1 - \theta/\sqrt{n}} \right) d\left( \frac{A^n(s) - U^n(s)}{n} \right). \quad (8)$$

Because $U^n/n \to 0$ (which follows because $\hat{U}^n \to \hat{U}$ as $n \to \infty$) and $A^n(n)/n \to s$ as $n \to \infty$, passing to the limit in (8) formally gives $\hat{\xi}^n \to \hat{\xi} \quad \text{as } n \to \infty$, where

$$\hat{\xi}(t) = RB(t) + \nu\theta t - \left[ R\hat{U}(t) + \int_0^t h(\hat{W}(s))ds \right], \quad t \geq 0.$$  

Then, taking the expectations of both sides gives the following upper bound on the expected cumulative profits up to time $t$:

$$E[\hat{\xi}(t)] = \nu\theta t - E\left[ R\hat{U}(t) + \int_0^t h(\hat{W}(s))ds \right], \quad t \geq 0.$$  

Clearly, this upper bound is policy dependent. Therefore, in the next section we first maximize this upper bound over the admission control policy $\hat{U}$ by considering the associated singular control problem. Then, we set the capacity imbalance term $\theta$ optimally, which yields the desired asymptotic upper bound.

5. Singular Control Problem and the Asymptotic Performance Bound

In this section, we maximize the asymptotic upper bound developed in §4 by first considering the following singular control problem, whose solution provides an optimal admission control policy. Then, we choose the optimal capacity imbalance term $\theta$, which yields the desired asymptotic upper bound on system performance. The singular control problem involves choosing nondecreasing processes $\hat{L}, \hat{U}$ adapted to $B$ so as to

minimize $\limsup_{t \to \infty} \frac{1}{t} E \left[ R\hat{U}(t) + \int_0^t h(\hat{W}(s))ds \right] \quad (9)$

subject to

$$\hat{W}(t) = B(t) + \theta t + \hat{L}(t) - \hat{U}(t), \quad t \geq 0,$$

$$\hat{W}(t) \geq 0, \quad t \geq 0,$$

$$\hat{\hat{L}}, \hat{\hat{U}} \text{ are nondecreasing with } \hat{\hat{L}}(0) = \hat{\hat{U}}(0) = 0. \quad (12)$$

Our analysis assumes only that $h(\cdot)$ is continuous and strictly increasing with $h(0) = 0$ and $\lim_{x \to \infty} h(x) = \infty$. In particular, we are not making any convexity or differentiability assumptions on $h(\cdot)$. Therefore, the analysis of
subject to the boundary conditions

\[ f'(0) = 0 \quad \text{and} \quad f'(x) = R \quad \text{for } x \geq w(\theta). \]  

Here one interprets \( \gamma(\theta) \) as a guess at the minimum average cost, and \( w(\theta) \) corresponds to an upper reflecting barrier to be imposed on the workload process. The unknown function \( f \) is often called the relative value function in average cost dynamic programming. It should be emphasized that the Bellman equation is introduced primarily to motivate our solution approach; the properties of the Bellman equation that we require will be proved from first principles (cf. Proposition 2 and Theorem 2).

Because the Bellman Equation (13)–(14) does not involve the function \( f \) itself, it is really a first-order equation. Therefore, defining \( \mathcal{C}[0, \infty) \) as the space of continuously differentiable functions and setting \( v(x) = f'(x) \) for \( x \geq 0 \), one can equivalently state the Bellman equation as follows. Choose a nondecreasing function \( v \in \mathcal{C}[0, \infty) \) and constants \( \gamma(\theta) > 0 \) and \( w(\theta) > 0 \), which jointly satisfy

\[ \gamma(\theta) = \frac{1}{2} \sigma^2 v'(x) + \theta v(x) + h(x) \quad \text{for } x \in [0, w(\theta)] \]  

subject to the boundary conditions

\[ v(0) = 0 \quad \text{and} \quad v(x) = R \quad \text{for } x \geq w(\theta). \]

The following definitions are needed to facilitate our solution of the Bellman equation:

\[ \bar{\gamma}_\theta = \begin{cases} \int_0^\infty \left( -\frac{2\theta}{\sigma^2} \right) \exp \left( \frac{2\theta}{\sigma^2 y} \right) h(y) \, dy & \text{if } \theta > 0, \\ \infty & \text{otherwise}, \end{cases} \]

and

\[ g(x) = h(x) \exp \left( \frac{2\theta}{\sigma^2 x} \right) - \int_0^x \frac{2\theta}{\sigma^2 y} \exp \left( \frac{2\theta}{\sigma^2 y} \right) h(y) \, dy, \]

\[ x \geq 0. \]

The following proposition characterizes the key properties of \( g \).

**Proposition 2.** \( g \) is strictly increasing on \([0, \infty)\) with \( g(0) = 0 \) and \( \lim_{x \to \infty} g(x) = \bar{\gamma}_\theta \).

It follows immediately from Proposition 2 that the inverse of the function \( g \) is well defined. Denote that inverse by \( g^{-1} \), which is strictly increasing on its domain \([0, \bar{\gamma}_\theta]\) with \( g^{-1}(0) = 0 \) and \( \lim_{y \to \bar{\gamma}_\theta} g^{-1}(y) = \infty \). To facilitate our analysis, we next consider the following initial value problem IVP(\( \gamma \)) for \( \gamma \in [0, \bar{\gamma}_\theta] \):

\[ \gamma = \frac{1}{2} \sigma^2 v'_\gamma(x) + \theta v_\gamma(x) + h(x) \quad \text{for } x \geq 0, \]  

\[ v_\gamma(0) = 0. \]

The following proposition provides the solution to the initial value problem (17)–(18).

**Proposition 3.** For \( x \geq 0 \) and \( 0 \leq \gamma < \bar{\gamma}_\theta \), one has that

\[ v_\gamma(x) = \begin{cases} \frac{\gamma}{\theta} \left[ 1 - \exp \left( -\frac{2\theta}{\sigma^2 x} \right) \right] - \frac{2}{\sigma^2} \exp \left( \frac{2\theta}{\sigma^2 x} \right) \int_0^x \exp \left( \frac{2\theta}{\sigma^2 y} \right) h(y) \, dy & \text{if } \theta \neq 0, \\ 2\gamma x - \frac{2}{\sigma^2} \int_0^x h(y) \, dy & \text{if } \theta = 0. \end{cases} \]

Define

\[ \psi(\gamma) = \sup_{x \geq 0} v_\gamma(x) \quad \text{for } 0 \leq \gamma < \bar{\gamma}_\theta. \]

Propositions 4 and 5 characterize the essential properties of \( \psi \).

**Proposition 4.** For \( 0 \leq \gamma < \bar{\gamma}_\theta \), \( v_\gamma \) achieves its maximum \( \psi(\gamma) \), and \( g^{-1}(\gamma) \) is the unique maximizer. That is, \( \psi(\gamma) = v_\gamma(g^{-1}(\gamma)) \). Moreover, \( v_\gamma \) increases strictly to its maximum.

**Proposition 5.** \( \psi \) is strictly increasing and continuous on its domain \([0, \bar{\gamma}_\theta]\) with \( \psi(0) = 0 \) and \( \lim_{\gamma \to \bar{\gamma}_\theta} \psi(\gamma) = \infty \).

It follows immediately from Proposition 5 that \( \psi \) is invertible. Denoting that inverse by \( \psi^{-1} \), it also follows from Proposition 5 that \( \psi^{-1} \) is continuous, strictly increasing on \([0, \infty)\) with \( \psi^{-1}(0) = 0 \) and \( \lim_{\gamma \to \infty} \psi^{-1}(\gamma) = \bar{\gamma}_\theta \).

The following proposition provides an explicit solution to the Bellman equation (15)–(16).

**Proposition 6.** Let \( \gamma(\theta) = \psi^{-1}(\theta) \), \( w(\theta) = g^{-1}(\gamma(\theta)) \), and

\[ v(x) = \begin{cases} v_{\gamma(\theta)}(x) & \text{if } x \leq w(\theta), \\ R & \text{if } x > w(\theta). \end{cases} \]

Then, \((v, \gamma(\theta), w(\theta))\) is the (unique) solution of the Bellman equation.

Then, defining

\[ f(x) = \int_0^x v(y) \, dy \quad \text{for } x \geq 0, \]

the following corollary provides a solution to the Bellman equation (13)–(14).
Corollary 1. The triple \((f, \gamma(\theta), w(\theta))\) solves the Bellman equation (13)–(14). Moreover, the function \(f\) is convex (and unique up to an additive constant).

The candidate policy for optimality imposes reflecting barriers for the workload process at zero and \(w(\theta)\) by appropriately choosing the controls \(L^*\) and \(U^*\). In particular, the controls \(L^*\) and \(U^*\) are continuous, and they increase only at zero and \(w(\theta)\), respectively. The evolution of the workload process \(W^*\) under the candidate policy \((L^*, U^*)\) can be described as follows:

\[
W^*(t) = B(t) + \theta t + L^*(t) - U^*(t), \quad t \geq 0.
\]

Moreover, the controls \(L^*\) and \(U^*\) and the corresponding workload process \(W^*\) jointly satisfy the following.

\[
W^*(t) \in [0, w(\theta)], \quad t \geq 0,
\]

\[
\int_0^t W^*(s) dL^*(s) = \int_0^t (w(\theta) - W^*(s)) dU^*(s) = 0, \quad t \geq 0,
\]

\(L^*\) and \(U^*\) are continuous and nondecreasing with \(L^*(0) = U^*(0) = 0\).

The following theorem establishes the optimality of the candidate policy \((L^*, U^*)\).

Theorem 2. For \(-\infty < \theta < \infty\), the candidate policy \((L^*, U^*)\) corresponding to reflecting barriers at zero and \(w(\theta)\) is optimal for the singular control problem (9)–(12), and its long-run average cost is \(\gamma(\theta)\).

The following proposition is needed to characterize the desired asymptotic upper bound.

Proposition 7. \(\gamma(\theta)\) is nonnegative, nondecreasing, and Lipschitz continuous in \(\theta\) with Lipschitz constant \(R\). We also have that

\[
\lim_{|\theta| \to \infty} (\nu \theta - \gamma(\theta)) = -\infty.
\]

Moreover, if \(h(\cdot)\) is (strictly) convex, then \(\gamma(\cdot)\) is also (strictly) convex.

The following corollary is immediate from Proposition 7.

Corollary 2. The set \(\arg \max \{\nu \theta - \gamma(\theta)\}\) is nonempty (and convex if \(h(\cdot)\) is convex).

The next result establishes the desired asymptotic bound and is proved in §B.2 of the online appendix.

Theorem 3. For any sequence of admissible policies \(\{\nu_n, n \geq 1\}\), we have that

\[
\lim_{t \to \infty} \lim_{n \to \infty} \frac{1}{T} \mathbb{E}[\hat{I}^T_{\nu_n}(T)] \leq \max_{\theta \in \Theta} (\nu \theta - \gamma(\theta)).
\]

For convex cost functions, the asymptotic upper bound corresponds to the asymptotically optimal cost because \(c(\cdot) = h(\cdot)\). For nonconvex cost functions, if we can asymptotically approximate the performance of the upper bound, then such a policy should be asymptotically optimal. In the following section we propose asymptotically optimal policies that are based on this intuition.

6. Proposed Policies

In this section, we introduce our discrete-review framework and describe the proposed policies in that framework. As argued in §4, the optimal system capacity should be of the form \(\mu^* = n - \theta \sqrt{n}\), which reduces the optimal capacity decision to choosing the capacity imbalance term \(\theta\). Fixing \(\theta^* \in \arg \max_{\theta \in \Theta} (\nu \theta - \gamma(\theta))\), we let

\[
\mu^* = n - \theta^* \sqrt{n} \quad \text{for} \quad n \geq 1.
\]

Then, we choose a review-period length (or a production-shift length) \(\kappa^*\) for each system. In particular, we let

\[
\kappa^* = z_1 \left[ \frac{n^{1-\alpha}}{\mu^*} \right] \quad \text{for} \quad n \geq 1,
\]

where \(1/2 < \alpha < 1\) and \(z_1 > 0\) is a fixed integer.

In the \(n\)th system, the system manager reviews the system status at times \(a_n = kn^\alpha\) for \(k = 0, 1, 2, \ldots\), and chooses how many customers to admit during the current period (which will be processed starting only in the next period). As the system manager admits customers, she quotes them lead times (and the corresponding prices) and decides the order (but not the exact timing) in which they will be processed during the upcoming periods so that the quoted lead times are respected. This results in a detailed production plan for the upcoming production shift so that the server can just process jobs as prescribed by this plan. The production plan for each shift concerns processing of only the jobs that are present at the beginning of the shift.

Given our choice of \(\kappa^*\), the server can process \(z_1 \left[ n^{1-\alpha} \right] \) jobs in each shift, and \(\kappa^*\) is of order \(1/n^\alpha\), which ensures that the system manager reviews the system status sufficiently frequently provided \(\alpha \in (1/2, 1)\). Indeed, our choice of \(\kappa^*\) is in line with the orders of magnitude of the length of review periods used in the earlier work on discrete-review policies for controlling stochastic networks; cf. Harrison (1996), Maglaras (2000), Ata and Kumar (2005), and Plambeck and Ward (2006).

Given the capacity choices and the discrete-review framework, we next describe the admission control, lead-time quotation, and sequencing decisions. As one would expect, different shapes of the disutility function for delay give rise to different control policies. Thus, we next describe the proposed control policy in each of the three cases. However, in all three cases the admission control policy has the same form, which is given below. This is a consequence of the lower bound analysis. We consider the policies for some large arrival rate \(\lambda = n\) and argue intuitively as to why the proposed policy should perform well. We support this intuition with a proof of asymptotic optimality of the proposed policies as the arrival rate gets large.

6.1. Admission Control Decisions

The admission control policy specifies the desired number of customer orders the system manager wishes to admit
over each review period. This is really an upper bound on the number of customer orders that can join the system in each review period, because when those customers arrive at the system, they are quoted a (nonzero) price and a lead time, and if the lead time is too large, i.e., \( c^n(\tau) > R \), then they may not join the system.

More specifically, we propose the following admission control policy: At the beginning of each review period \( t_k \) for \( k = 0, 1, 2, \ldots \), the system manager observes the total number of jobs \( Q^i(t_k) \) and wishes to admit \( \lceil w(\theta) \sqrt{n} + \mu \kappa^x \rceil \) in the current period (i.e., during \( (t_k, t_{k+1}] \)), where \( w(\theta) \) is from Proposition 6. If the number of arrivals is less than \( \lceil w(\theta) \sqrt{n} + \mu \kappa^x - Q^i(t_k) \rceil \), then the system manager admits all arrivals during the current period. Note that this admission control policy ensures that \( Q^i(t) \) is less than or equal to \( \lceil w(\theta) \sqrt{n} + \mu \kappa^x \rceil \) at all times. Moreover, for \( n \) large enough (i.e., \( w(\theta) \sqrt{n} > \mu \kappa^x \)), under the proposed policy the number of jobs in the system \( Q^i(t_k) \) at each review point is less than or equal to \( \lceil w(\theta) \sqrt{n} \rceil \) and at least \( \mu \kappa^x \) jobs are admitted during every period (provided there are sufficiently many arrivals).

Note that, although the form of the admission control policy is the same for all policies, the actual value of the parameter \( w(\theta) \) depends on \( h(-) \), the convex hull of the delay cost function. Once the admission control rule has been chosen, it remains to specify the lead-time quotation and scheduling rules. This is done separately below for each of the three forms for the cost function. Although the proposed policy in the convex case is rather straightforward, it is more involved in the other two cases. In those cases, the main idea is to create two virtual service classes and to give priority to class 1 while providing a small but positive amount of service capacity to the low-priority class. The following subsections elaborate on the intuition behind this approach and why it leads to near-optimal policies.

### 6.2. Lead-Time Quotation and Sequencing Decisions

To facilitate our description of the proposed policies, consider Figure 2. The left panel of Figure 2 displays a concave disutility of delay function and its convex hull, and the right panel displays a convex-concave disutility of delay function and its convex hull; in each case, the solid curve depicts the disutility of delay function and the dashed curve corresponds to its convex hull.

Our analysis assumes that \( \lim_{n \to \infty} c(x) = c > 0 \). We also assume that \( c'(0) < \infty \). Then, as can also be seen from the left panel of Figure 2, the convex hull \( h(-) \) of a concave disutility of delay function \( c(-) \) is given by

\[
h(x) = cx, \quad x \geq 0.
\]

Similarly, as can be seen from the right panel of Figure 2, the convex hull \( h(-) \) of a convex-concave function \( c(-) \) is given by

\[
h(x) = \begin{cases} 
  c(x) & \text{if } x \leq x, \\
  c(x) + (x - \varepsilon)c & \text{otherwise},
\end{cases}
\]

where \( \varepsilon = \min\{x \geq 0: c'(x) \geq c\} \). On the other hand, the convex hull of a convex disutility of delay function is clearly itself.

The idea behind the lead-time quotation and sequencing policies we propose is to ensure that, when the workload is \( \nu \), a delay cost rate of \( h(\nu) \) is incurred. This can easily be done in the convex case by sequencing the jobs on an FCFS basis and quoting lead times accordingly, because the convex hull of a convex function is itself. In contrast, the system manager must use a fundamentally different approach in the other two cases. In these two cases, the proposed lead-time quotation and scheduling policy is specified by a high-priority or class 1 service capacity \( \eta_1 \) and a low-priority or class 2 service capacity \( \eta_2 \) (\( \eta_2^i \) denotes the number of class \( i \) jobs that can be served in a production shift), where \( \eta_2^i = z_2 [n^\beta] \) (with \( (1 - \alpha)/2 < \beta < (1 - \alpha) \) and \( z_2 \) a positive integer) is small enough that \( \eta_2^i = o(n^{1-\alpha}) \) as \( n \) gets large, and \( \eta_1^i = \mu \kappa, - \eta_2^i \). Because \( \eta_2 \) is just required to be “small enough” its value may be fine-tuned via simulation. Next, we lay out the details of the proposed sequencing and lead-time quotation policy in each of the three cases.

#### Proposed Policy for the Case of Convex Disutility of Delay

For convex disutilities, we propose that a job admitted to the system at time \( t \) is quoted a lead time of \( \lceil Q^i(t) / \mu \kappa \rceil + 1 \) periods, where \( Q^i(t) \) denotes the number of jobs in the system; and the admitted jobs are served on an FCFS basis. A job admitted during period \( k \) can be processed only in period \( k + 1 \) or later, which is why the lowest possible lead time quoted is one period. As suggested by Theorem 1, this is actually the optimal lead-time quotation policy in our discrete-review framework. Indeed, Theorem 4 proves that the sequence of policies we propose is asymptotically optimal in the convex case.

#### Proposed Policy for the Case of Concave Disutility of Delay

Given the parameters \( \eta_1^i \) and \( \eta_2^i \) from above, the proposed policy can be described as follows. Create two virtual classes indexed by \( i = 1, 2 \); and let \( Q^i(t) \) denote the number of class \( i \) jobs in the system at time \( t \). In every period, route the first \( \eta_1^i \) admitted jobs to class 1, quoting them the lead time of one. Route the rest of the arrivals to class 2, quoting an arrival at time \( t \) the lead time of
To elaborate on the intuition behind this policy, recall that the convex hull \( h(\cdot) \) of a concave delay cost function \( c(\cdot) \) is a linear function with slope \( c \). The linear form of the convex hull function suggests that under an asymptotically optimal policy if a customer is quoted a lead time that is larger than the minimal allowable lead time, his marginal cost of delay should equal \( c \), which in turn suggests that the system manager should quote a large lead time whenever she quotes one larger than the minimum allowable lead time. Although this penalizes those customers, it also frees up some capacity in the short term for newly arriving jobs who may then be quoted the minimal allowable lead time of one. Indeed, under the proposed policy, the majority \((\eta^i_n)\) of customers receive a lead time of one. However, a small fraction of customers \((\eta^2_n)\) are kept as a hedge against demand uncertainty and are quoted very long lead times. Hence, one would expect this policy to perform well. Indeed, Theorem 4 establishes its asymptotic optimality in the large-capacity asymptotic regime.

### Proposed Policy for the Case of Convex-Concave Disutility of Delay

In this case, the proposed lead-time quotation and scheduling policy is also specified by a high-priority or class 1 service capacity \( \eta^1_n \) and a low-priority or class 2 service capacity \( \eta^2_n \). The proposed policy can be described in terms of these parameters as follows.

As in the concave case, create two virtual classes indexed by \( i = 1, 2; \) and let \( Q^i_n(t) \) denote the number of class \( i \) jobs in the system at time \( t \). To facilitate our description of the proposed sequencing and lead-time quotation rules, it is helpful to imagine that we have two server pools, each of which consists of fully flexible servers that can process either class. Server pool \( i \) serves class \( i \) and has service capacity of \( \eta^i_n \) in each period. (This analog system merely corresponds to our server splitting its time between the two classes over each production shift so that class \( i \) receives a service capacity of \( \eta^i_n \).) Each server pool gives priority to its own customers, but collectively the service policy is nondelaying in the sense that neither server idles until all jobs that were present at the beginning of the period are processed. (Recall that the servers do not work on the jobs that arrive during the current period.) Then, a customer routed to class \( i \) at time \( t \) is quoted a lead time of \( \tau^i_n(t) = \lceil Q^i_n(t)/\eta^i_n \rceil + 1 \) for \( i = 1, 2 \). As for the routing policy, at each review point \( t_k \) \((k = 0, 1, \ldots,)\), the system manager observes the system status. Then, she routes the first \( \eta^i_n \) admitted jobs during period \( k \) to class 1 and routes the next \( \eta^2_n \) jobs to class 2. Then, the next \( \lceil \sqrt{n} \rceil - Q^2_1(t_k) \) jobs are again routed to class 1. Any further admitted jobs in that period are routed to class 2. Note that this routing policy ensures that the number of class 1 jobs \( Q^1_n(t) \) is less than or equal to \( \lceil \sqrt{n} \rceil + \eta^1_n \) at all times. Moreover, for \( n \) large enough, i.e., \( \lceil \sqrt{n} \rceil > \eta^1_n \), under the proposed policy the number of class 1 jobs \( Q^1_n(t_k) \) at each review point \( t_k \) is less than or equal to \( \lceil \sqrt{n} \rceil \).

Once again, the proposed policy strives to achieve the performance under the convex hull function asymptotically. Recall that the convex hull function \( h(\cdot) \) coincides with the delay cost function \( c(\cdot) \) if the (scaled) workload is less than or equal to \( x \), and is a line with slope \( c \) otherwise (cf. the right panel of Figure 2). The structure of the convex hull function \( h(\cdot) \) suggests that the system manager should use the FCFS sequencing rule and quote the minimal possible lead time whenever the backlog of work in the system is sufficiently low, that is, whenever the (scaled) workload is less than \( x \). On the other hand, when the backlog of work is large the convex hull function is linear with slope \( c \), suggesting that under an asymptotically optimal policy the marginal cost of waiting should be close to \( c \) for those customers who are quoted lead times beyond the range where the marginal cost of delay is larger than \( c \).

The above observation then advocates quoting either moderate delays, which are less than or equal to \( \sqrt{n} \), or large delays when the backlog of work is large. The proposed policy accomplishes this as follows. When the backlog of work is small to moderate, the arrivals are routed to both classes, served on an FCFS basis, and quoted lead times accordingly. During such periods, because the number of jobs routed to class 2 is less than or equal to the capacity \( \eta^2_n \) of the second server pool, no backlog accumulates in class 2 and all backlog is kept in class 1. In contrast, when the backlog in the system is large, the system manager puts a cap on the number of class 1 jobs in the system, routing all jobs to class 2 beyond that cap, resulting in a large backlog in class 2, which in turn translates into large delays for those jobs because the capacity of the second server pool is small. In other words, whenever the (scaled) backlog in the system exceeds \( x \), class 1 backlog is \( x \) and the remaining backlog is kept in class 2. If the (scaled) backlog is less than \( x \), then all backlog is kept in class 1. This is precisely a state-space collapse result under the proposed policy (cf. §B.3.4 of the online appendix), and it lets the system manager achieve the delay cost rates corresponding to \( h(\cdot) \) asymptotically. The following theorem states the asymptotic optimality results and is proved in §B.3 of the online appendix.

**Theorem 4.** For each of the three cases, the sequence of policies proposed in this section is asymptotically optimal. In particular, they achieve the asymptotic upper bound provided in Theorem 3.

### 7. Numerical Study

In this section, we present two numerical examples to illustrate the effectiveness of the proposed policies. Ideally, one
should compare the proposed policies against the optimal ones to assess their performance, but computing the optimal policies in our setting appears to be virtually impossible for the cases of concave and convex-concave disutility of delay. Given that we study a system with homogeneous customers, an obvious benchmark policy is the one that uses the FCFS sequencing policy and quotes lead times accordingly, for which we also find the optimal capacity and the admission control threshold, by solving the singular control problem of §5 for the appropriate cost-of-delay function. Because the proposed policy reduces to the benchmark policy in the case of convex disutility of delay, we will make the comparison only for the cases of concave and convex-concave disutility of delay.

Our goal is not to evaluate the effectiveness of our policies relative to the benchmark for all possible shapes of cost function and all possible ranges of parameter values. Indeed, we believe that such a study makes little sense without industry data to guide the values and functions. Instead, we seek to see if it is indeed possible for the proposed policies to yield significant cost savings and to provide insight into when the proposed policies may result in such savings.

In our simulation study, we assume for simplicity that the customers arrive according to a Poisson process and that there is a per-customer reward \( R = 8 \) from service and a cost of capacity per unit of capacity per unit of time of \( \nu = 5 \). We assume that the arrival rate is \( \lambda^0 = 1,000 \); i.e., we set the system parameter \( n = 1,000 \). First, we focus on the case of concave disutility of delay, where the (limiting) disutility of delay function is given by

\[ c(x) = \sqrt{x} + 0.1x \quad \text{for } x \geq 0. \]

Then, using the convex hull of this, we find \( \theta^* = -0.02 \) (cf. §5), which then gives the optimal capacity \( \mu^0 = 1,000.63 \). We also set \( z_1 = 4 \) and \( \alpha = 0.6 \), which gives \( \kappa^0 = 0.063 \). Recall that \( \kappa^0 \) was interpreted as the length of a production shift. To be concrete, given that the length of a production shift is frequently around eight hours, an absolute unit of time in our model corresponds to 3.16 weeks (assuming one shift per day and five working days per week). Then, \( n = \lambda^0 = 1,000 \) corresponds to receiving 315 requests per week on average.

In our benchmark study, the optimal capacity and the admission control threshold are determined by the analysis presented in §5 (with the delay cost function \( c(\cdot) \) in place of \( h(\cdot) \)), which results in the admission control threshold value of 221. That is, there can be at most 221 jobs in the system; the admitted customers are served on an FCFS basis and are quoted lead times accordingly. A histogram of delays quoted under the benchmark policy is given in Figure 3, which shows that the delays under the benchmark policy are indeed moderate. The 95% confidence interval\(^2\) for the average (scaled) profit (normalized by subtracting the upper bound \( (R - \nu)nT \)) associated with our benchmark policy is \(-2.82 \pm 0.05\).

The admission control threshold for the proposed policy is 365. The remaining parameters of the proposed policy \( \eta^c_1 = 59 \) and \( \eta^c_2 = 5 \) (letting \( \beta = 0.21 \) and \( z_2 = 1 \)), whereas the 95% confidence interval for the average (scaled) profits is \(-2.00 \pm 0.07\), corresponding to an improvement of \( 29.05 \pm 1.61\% \) over the benchmark policy. To gain insight into why the proposed policy performs well, next consider the delays quoted under the proposed policy versus those under the benchmark policy. Under the proposed policy, 90.43% of the customers are routed to class 1 and 9.57% of them are routed to class 2. Clearly, under the proposed policy, class 1 customers experience the minimum possible delay, and the histogram of lead times quoted to class 2 customers is given in Figure 4. Essentially, the proposed policy provides excellent service to most customers while severely “punishing” a small fraction of them by quoting them large lead times, which is intuitive given the concave shape of the disutility of delay. In contrast, under the benchmark policy, all customers experience moderate delays, which is worse, again because of the concave shape of the disutility of the delay function.

We conclude our discussion of the concave case by the following sensitivity analysis. Table 1 summarizes the results. Each entry of the table reports a 95% confidence interval for the percentage improvement of the proposed policy over the benchmark policy. Each column of Table 1 corresponds to a particular system size, i.e., arrival rate, and each row corresponds to a particular value of the cost.

---

**Table 1:** Summary of percentage improvements for the proposed policy over the benchmark policy for the case of concave disutility of delay.

<table>
<thead>
<tr>
<th>Arrival Rate</th>
<th>Class 1 Improvement</th>
<th>Class 2 Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>97.5%</td>
<td>90.43%</td>
</tr>
<tr>
<td>2,000</td>
<td>98.8%</td>
<td>91.32%</td>
</tr>
<tr>
<td>3,000</td>
<td>99.0%</td>
<td>91.57%</td>
</tr>
</tbody>
</table>

---

**Figure 3.** A histogram of delays quoted under the benchmark policy in the case of concave disutility of delay.

**Figure 4.** A histogram of delays quoted to class 2 under the proposed policy in the case of concave disutility of delay.
of capital $\nu$. Recall that we arrived at the singular control problem and the proposed policies through a limiting argument as the system size grows. Thus, one would expect that the proposed policy will do better as the system grows, which is indeed the case as can be seen from Table 1. Similarly, as the capacity cost $\nu$ increases, the importance of scheduling decisions increases. Thus, the savings from the proposed policy increases. Indeed, in all the cases considered, the proposed policy uses less capacity, admits more customers, and yields higher profits. Table 2 shows the rejection rates under the benchmark policy and the proposed policy. (For brevity, we display only the result for $n = 1,000$; the results in the other cases are quite similar.)

Next, we consider the convex-concave case, where the (limiting) cost of delay function is

$$c(x) = \begin{cases} \frac{x^2}{5} & \text{if } x \leq 3, \\ \sqrt{x} + 0.1x - 0.232 & \text{otherwise,} \end{cases}$$

which is continuous and chosen so as to ensure that both the convex and the concave ranges are relevant under the benchmark policy. This is essential to assess the value of the proposed policy over the benchmark policy; if only the convex portion of the delay cost function is relevant, then the benchmark policy is indeed optimal (cf. Theorem 1). The remaining system data are the same as in the concave case. The optimal capacity associated for the proposed policy is 1,000.63, which yields $\kappa^n = 0.063$. The admission control parameter of the proposed policy is 365. The optimal capacity and the admission control threshold for the benchmark policy are determined by the analysis of §5 with the delay cost function $c(\cdot)$ in place of $h(\cdot)$, which results in the optimal capacity of $\mu^n = 1,001.58$ the admission control threshold of 182.

The histogram of delays quoted under the benchmark policy is given in Figure 5, which shows that the delays are indeed moderate and that both the convex and the concave parts of the delay cost function are relevant in the benchmark case. The 95% confidence interval for the average scaled profit with our benchmark policy is $-3.72 \pm 0.03$. On the other hand, the parameters for the proposed policy are as follows: The threshold for routing is 67, while $\eta = 59$ and $s\mu = 5$, and the 95% confidence interval for the scaled profit under the proposed policy is $-1.24 \pm 0.04$, corresponding to an improvement of 66.76 ± 1.60% over the benchmark policy.

Table 2. Percentage of arrivals not joining the system in the concave case.

<table>
<thead>
<tr>
<th>$\nu$</th>
<th>FCFS</th>
<th>The proposed policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.02 ± 0.00</td>
<td>0.02 ± 0.00</td>
</tr>
<tr>
<td>5</td>
<td>0.20 ± 0.01</td>
<td>0.13 ± 0.01</td>
</tr>
<tr>
<td>7</td>
<td>0.34 ± 0.02</td>
<td>0.17 ± 0.02</td>
</tr>
</tbody>
</table>

To gain insight, consider the histograms of quoted lead times under the benchmark policy and the proposed policy (cf. Figures 5 and 6). The lead times quoted under the benchmark policy are similar to those in the concave case. However, the lead-time histograms under the proposed policy show differences from the concave case. In the convex-concave case, an even smaller fraction (7.62%) of the customers are routed to class 2. These customers are quoted very large delays as in the concave case. Although class 1 delays under the proposed policy are small, they may be larger than one reflecting the convex-concave shape of the disutility of delay function. In essence, in the range where the disutility of delay is convex the proposed policy behaves like the benchmark policy, but beyond a certain threshold (that is, when the number of customers in the

Figure 5. A histogram of delays quoted under the benchmark policy in the case of convex-concave disutility of delay.

Figure 6. A histogram of delays quoted under the proposed policy in the case of convex-concave disutility of delay.

Note. The top panel shows the histogram for class 1, and the bottom panel shows that for class 2.
system grows and as the cost of capital increases. Moreover, the gains increase as the system size percentage) improvement depends on the parameters of the concave cases, we conclude that whereas the exact (per-
in the other cases are quite similar.)

We considered a system where customers have homoge-
nous but general (nonlinear) disutility for delay. For the case of convex (or linear) disutility, an FCFS lead-time quotation policy was shown to be optimal. This is comforting given how widely FCFS policies are used in practice. However, for concave or convex-concave disutilities, large improvements were shown to be possible over FCFS. The policies used a two-class scheduling system that was both highly intuitive and readily implementable. The key insight is to provide high-quality service to a majority of customers but to artificially segment a second low class who are quoted a long lead time and kept as a hedge against demand uncertainty.

We also considered the problem of capacity setting in this environment. It was shown that the system manager should optimally set capacity close to the arrival rate so that heavy-traffic conditions arise. Under these conditions, we showed our proposed policies to be asymptotically optimal.

We considered a monopolist who was able to extract all surplus from her customers so that customers were indifferent to a long lead time at a low cost or a short lead time at a higher price. One could instead consider the problem of maximizing system welfare. Here the system seeks to maximize total benefits from acceptance and service of customers minus delay costs. Maximizing total benefits minus costs may also be viewed as a system where the lead-time cost is an implicit cost (e.g., goodwill costs plus competitive losses) incurred by the firm for customer lead times. However, in such a setting the customers are no longer indifferent between the lead times quoted and hence such a model would need to incorporate the likely strategic behavior of customers. This is left as the subject of future research.

When customers are heterogenous in value to the firm but not lead-time tolerance, then the segmentation suggested here could easily be done on that basis rather than the random assignments currently assumed. This would be expected to result in even greater savings than those found to be possible here based on lead-time cost savings alone. Future research should consider more detailed consumer modeling. In particular, to isolate the effects of the shape of the lead-time disutility function, we considered neither heterogenous customers nor competition. Such extensions should be the subject of future work.

8. Conclusions

We considered a system where customers have homogenous but general (nonlinear) disutility for delay. For the case of convex (or linear) disutility, an FCFS lead-time quotation policy was shown to be optimal. This is comforting given how widely FCFS policies are used in practice. However, for concave or convex-concave disutilities, large improvements were shown to be possible over FCFS. The policies used a two-class scheduling system that was both highly intuitive and readily implementable. The key insight is to provide high-quality service to a majority of customers but to artificially segment a second low class who

9. Electronic Companion

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

Endnotes

1. For ease of exposition, we will keep the genders assigned in this sentence throughout the paper.
2. Alternatively, one can imagine that the system manager charges price \( p > R \) whenever she wants to induce a customer not to join the system.
3. Payments are based on quoted lead times, not actual lead times.
4. Note that this specific form of the cost of delay is not used in the proof of Proposition 1, which is valid for any nondecreasing cost function with \( \lim_{x \to \infty} c^n(x) = \infty \) for \( n \geq 1 \).
5. Each confidence interval referred to in this section is based on 80 simulation runs, each of which is for roughly 50 million arrivals to the system. Moreover, the first 20% of each simulation run is deleted to eliminate transient effects.

**Acknowledgments**

The authors thank Mustafa Tongarlak for performing much of the numerical study. They thank the anonymous associate editor and three reviewers whose comments resulted in a significantly improved paper. This research was supported in part by National Science Foundation grant DMI-0245382. The former title of this paper was “Dynamic Lead-Time Quotation Under General Customer Utilities.”

**References**


Guo, P., P. Zipkin. 2007c. The effects of information on a queue with balking and phase-type service times. Working paper, Duke University, Durham, NC.


Kumar, S., R. S. Ramdahwa. 2008. Exploiting market size in service systems. Working paper, Graduate School of Business, Stanford University, Stanford, CA.


