PERFORMANCE ANALYSIS AND OPTIMIZATION OF ASSEMBLE-TO-ORDER SYSTEMS WITH RANDOM LEAD TIMES

JING-SHENG SONG

Graduate School of Management, University of California, Irvine, California 92697, jssong@uci.edu

DAVID D. YAO

Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027, yao@ieor.columbia.edu

(Received January 2000; revisions received January 2001, April 2001; accepted April 2001)

We study a single-product assembly system in which the final product is assembled to order whereas the components (subassemblies) are built to stock. Customer demand follows a Poisson process, and replenishment lead times for each component are independent and identically distributed random variables. For any given base-stock policy, the exact performance analysis reduces to the evaluation of a set of $M/G/\infty$ queues with a common arrival stream. We show that unlike the standard $M/G/\infty$ queueing system, lead time (service time) variability degrades performance in this assembly system. We also show that it is desirable to keep higher base-stock levels for components with longer mean lead times (and lower unit costs). We derive easy-to-compute performance bounds and use them as surrogates for the performance measures in several optimization problems that seek the best trade-off between inventory and customer service. Greedy-type algorithms are developed to solve the surrogate problems. Numerical examples indicate that these algorithms provide efficient solutions and valuable insights to the optimal inventory/service trade-off in the original problems.

1. INTRODUCTION

An assemble-to-order (ATO) system is an important business model in managing a wide-ranging class of supply chains. Perhaps the best way to understand and appreciate an ATO system is to consider the manufacturing and distribution of PCs (personal computers). A PC is a complex machine, built with hundreds of components. A PC company typically offers several lines of product, with each allowing dozens if not hundreds of "features" from which customers can select when placing an order—different combinations of CPU, memory, hard drive, and other components and peripherals (CD ROM, sound card, modem, monitor, keyboard, printer, etc). Whereas each of these components takes a substantial lead time to build, the time it takes to assemble all the components into a PC, following a specific customer order, takes virtually no time—provided all the components are available. Hence, managing the component inventory is of critical importance to the business: The stockout of any component will delay order fulfillment, whereas excess inventory could easily wipe out the firm's profit margin and diminish its competitive edge.

The objective of this paper is to address the optimal trade-off between inventory and service in an ATO system with m different components and a single end product. We assume that customer demand for the product arrives at the system following a stationary Poisson process. The replenishment lead times for each component are i.i.d. (independent and identically distributed) random variables. An independent base-stock (or one-for-one replenishment) policy is followed to control each component's inventory. We study two optimization problems to determine the optimal component base-stock levels. One problem is to minimize the expected number of backorders subject to an upper limit on the total component inventory investment. The other is to minimize the average component inventory subject to a fill-rate requirement for customer orders.

In the literature, studies of ATO systems mostly focus on base-stock control. This focus is appropriate in two regards. First, it is a well-studied policy with many familiar properties (e.g., Clark and Scarf 1960, Federgruen and Zipkin 1986, Glasserman 1997, Glasserman and Wang 1998, Rosling 1989). Second and more importantly, a base-stock policy provides a benchmark on how much inventory is needed to provide a certain service level. In this sense it sharpens the focus on the higher-level business issue of inventory/service trade-off, without getting into operational issues such as order sizes. (An exception is Song (2000), in which the (R, nQ) policy is considered for systems with constant lead times. However, the key point in that paper is that under certain general conditions, the analysis of the general reorder-point, order-quantity system reduces to that of the base-stock systems.)

Beyond this common focus on base-stock control, the studies differ quite substantially in the detailed modeling assumptions and approaches. One body of work considers periodic-review (discrete-time) models and assumes multivariate normal distributions for demand and constant lead times for component replenishment. See, for example,

Subject classifications: Inventory/production: multi-item; operating characteristics; stochastic models.
Area of review: Stochastic Models.

0030-364X/02/5005-0889 $05.00
1526-5463 electronic ISSN

Operations Research © 2002 INFORMS
Vol. 50, No. 5, September–October 2002, pp. 889–903

Agrawal and Cohen (2001) and Zhang (1997), and the references therein.

Another body of work, which is more closely related to this study, focuses on continuous-review models and assumes multivariate (compound) Poisson demand process. Several papers employ different component supply submodels than the one used in this paper. For systems with deterministic lead times, Song (1998, 2002) develops exact and approximate performance evaluation procedures that are computationally efficient. In Song et al. (1999) the supply system of each component is modeled as an independent production facility with a single exponential processor and a finite buffer, an $M/M/1/c$ queue. An exact performance analysis is carried out using matrix-geometric techniques. Glasserman and Wang (1998) model the supply system as a set of $M/G/1$ queues ($G/G/1$ queues for the single-product case) driven by a common demand stream, focusing on the delivery time and inventory trade-off. Based on the large deviations approach, a linear relationship between delivery time and inventory is established in the limiting sense of high fill rates. Wang (1999) further applies this asymptotic result in an optimization problem to minimize average inventory holding cost with a constraint on the order fill rate within a time window.

In contrast to modeling the component supply systems as single-server queues, here we use an infinite-server model; in particular, we treat the lead times for each component as i.i.d. (which includes the constant lead time model as a special case). Effectively, we are assuming an uncapacitated supply mechanism, or simply aggregating the queueing and processing delays of a replenishment order into a single piece, the lead time. This is appropriate when the supply sources are exogenous, and is a commonly used model in the inventory literature; see, e.g., Sherbrooke (1992) and Chapter 7 in Zipkin (2000).

Cheung and Hausman (1995) and Gallien and Wein (2001) also assume i.i.d. component lead times in the ATO context. However, they differ from our work in other aspects. For example, order synchronization is assumed in Gallien and Wein. That is, the replenishments of all components triggered by the same customer demand are later assembled into the same final product. Hence, effectively, there is only a single lead time, the one that corresponds to the longest component lead time. Cheung and Hausman (1995) use a combination of order synchronization and disaggregation in their analysis.

We assume no order synchronization in this paper, and our starting point is a direct, exact analysis.

Because each demand requires the simultaneous availability of several components, the correlation of demand at the component level is a central issue in analyzing ATO systems. As we shall see in Proposition 1, evaluating basic performance measures such as the back-order and the fill rates is a computationally hard problem, due to the "curse of dimensionality:" The underlying (joint) probability distribution from which the performance measures are computed involves $2^m - 1$ factors, with m being the number of components and each factor representing the status of a subset of components.

To overcome this difficulty, our approach is to (a) develop upper and lower bounds, which are easily computable, for the performance measures; (b) use the bounds as surrogates in the optimization problems; and (c) find effective solution procedures to solve the surrogate optimization problems. Numerical studies indicate that solutions to the surrogate problems via upper/lower bounds are in most cases optimal or near optimal (as verified by detailed simulation).

Some of our bounds are similar to those in Xu and Li (2000) in that they are also based on notions such as association and stochastic orders. The difference is that while Xu and Li is concerned with the comparison between different types of customer orders, our results here relate the computationally unwieldy system with random lead times to the tractable system with constant lead times, in terms of performance bounds. Some of our bounding techniques and the idea of using the bounds as surrogate objectives can be traced to Connors and Yao (1996), although the application context is quite different. The inventory/service trade-off, in the context of supply networks recently studied in Ettl et al. (2000) is quite similar to the stated objective of this paper; however, our focus on ATO systems and the detailed technical treatment here are new.

The rest of the paper is organized as follows. In §2 we present the model details and the main performance measures of interest and derive the joint distribution of outstanding component orders. In §3 we analyze the qualitative effect of lead-time variability and base-stock levels on system performance, and derive upper and lower bounds for the main performance measures. These bounds are used as surrogates in the optimization models in §4 and §5 which, respectively, minimize the number of back orders subject to an upper limit on inventory investment and minimize inventory subject to a fill-rate requirement. In both sections, we develop greedy-type algorithms to obtain solutions and, wherever applicable, prove their optimality. Numerical examples are illustrated in §6, where several observations on the optimal inventory/service trade-off are also discussed.

2. MODEL DETAILS AND PERFORMANCE MEASURES

Let $\mathcal{J} = \{1, 2, \ldots, m\}$ denote the set of all component indices. The end product consists of exactly one unit of each component in $\mathcal{J}$. (If the end product requires multiples of one component, we can simply redefine the unit of that component, and adjust accordingly the lead time for producing the unit.) Demand for the product arrives according to a Poisson process $\{A(t), t \geqslant 0\}$, with rate λ. Demands are filled on a first-come-first-served (FCFS) basis. If there is positive on-hand inventory for all components upon a demand arrival, the demand is filled immediately. In other words, we assume that the time to assemble the components

into the end product is negligible. On the other hand, if there is a stockout at one or more of the component inventories upon the demand arrival, then the demand is backlogged until the stockout components become available.

Let G_i be the common cumulative distribution of the replenishment lead times of component i, and let L_i denote the generic random variable with distribution G_i and mean $\mathsf{E}[L_i] = \ell_i$. Denote $\overline{G}_i = 1 - G_i$. Assume the lead times are independent among the components; that is, L_i is independent of L_j for any $i \neq j$.

We control the inventory of each component by an independent base-stock policy, with

$s_i :=$ the base-stock level for component i.

That is, at each demand arrival epoch, if the inventory position (i.e., the on-hand inventory plus on-order position minus backorders) of component i is less than s_i, then order up to s_i; otherwise, do not order.

Since the final product consists of a single unit of each component, the base-stock policy implies that every demand will trigger a replenishment order of one unit of each component, regardless of whether or not there is a stockout at the component inventory. Therefore, the outstanding orders for each component i at any time t, denoted $X_i(t)$, is equal to the number of jobs in service in an $M/G_i/\infty$ queue, for $i = 1, \ldots, m$. Note that the m queues are driven by a common Poisson arrival process $\{A(t)\}$, and hence are *not* independent.

For any given time t, the performance measures of interest are:

$I_i(t)$ = inventory on hand of component i at time $t = [s_i - X_i(t)]^+$;
$B_i(t)$ = number of backorders of component i at time $t = [X_i(t) - s_i]^+$;
$B(t)$ = number of backordered demand (for the final product) at time t;

where $x^+ := \max\{0, x\}$. Note that from the FCFS rule, we have

$$B(t) = \max_{1 \leqslant i \leqslant m} B_i(t) = \max_{1 \leqslant i \leqslant m} [X_i(t) - s_i]^+.$$

Let X_i, I_i, B_i, and B be the corresponding steady-state limits of the abovementioned random variables. Then,

$$I_i = [s_i - X_i]^+, \qquad B_i = [X_i - s_i]^+, \tag{1}$$

$$B = \max_{1 \leqslant i \leqslant m} B_i = \max_{1 \leqslant i \leqslant m} [X_i - s_i]^+. \tag{2}$$

We are also interested in the following order-fulfillment service measures:

f_i = fill rate of component i
= the probability of immediately satisfying a demand for component i,
f = order fill rate
= probability that a demand of the final product is filled immediately.

Due to the property that Poisson Arrivals See Time Average (e.g., Wolff 1989), we have

$$f_i = \mathsf{P}(I_i > 0) = \mathsf{P}(X_i < s_i),$$
$$f = \mathsf{P}(I_1 > 0, \ldots, I_m > 0) = \mathsf{P}(X_1 < s_1, \ldots, X_m < s_m). \tag{3}$$

From standard queuing results, the marginal distribution of X_i follows a Poisson distribution with mean $\lambda\ell_i = \lambda\mathsf{E}[L_i]$, which depends on the lead time distribution G_i only through its mean. This implies that the higher moments of the lead time (variance in particular) do *not* affect the component-based performance f_i, I_i, and B_i. This is no longer true, however, for the joint distribution of $(X_i, i = 1, \ldots, m)$, as we shall see later. The higher moments of lead times *do* affect product-based performance measures f and B.

Throughout the paper, we shall use $N(a)$ to denote a Poisson random variable with parameter (mean) a, along with the following notation:

$$p(n|a) := \mathsf{P}\,[N(a) = n] = \tfrac{a^n}{n!} e^{-a};$$
$$P(n|a) := \mathsf{P}\,[N(a) \leqslant n] = \textstyle\sum_{k=0}^{n} p(k|a);$$
$$\overline{P}(n|a) := \mathsf{P}\,[N(a) > n] = 1 - P(n|a).$$

Since $X_i \stackrel{d}{=} N(\lambda\ell_i)$, we can express the component-based performance measures as follows (see, e.g., Chapter 6 in Zipkin 2000):

$$f_i = \mathsf{P}(X_i < s_i) = \mathsf{P}(s_i - 1|\lambda\ell_i); \tag{4}$$

$$\mathsf{E}[B_i] = \lambda\ell_i - \sum_{n=0}^{s_i - 1} \overline{P}(n|\lambda\ell_i); \tag{5}$$

$$\mathsf{E}[I_i] = s_i - \mathsf{E}[X_i] + \mathsf{E}[B_i] = s_i - \lambda\ell_i + \mathsf{E}[B_i]. \tag{6}$$

Next, we study the joint distribution of the on-order positions $\{X_1(t), \ldots, X_m(t)\}$. For ease of exposition, we start with a system of two components, i.e., $m = 2$. Let $N_0(t)$ denote the number of those orders (jobs) that have arrived in $[0, t]$ and are still in service at both queues. Let $N_1(t)$ [resp. $N_2(t)$] denote the number of those orders that have arrived in $[0, t]$ and are still in service at Queue 1 [resp., Queue 2], but not at Queue 2 [resp., Queue 1]. Hence, at time t, there are

$$X_i(t) = N_0(t) + N_i(t) \tag{7}$$

jobs (outstanding orders) in queue i, $i = 1, 2$.

Consider a given $t > 0$. Suppose $A(t) = n$. Then, it is well known (e.g., Ross 1996) that the n (unordered) arrivals follow an i.i.d. uniform distribution in $[0, t]$, and that $(N_0(t), N_1(t), N_2(t))$ follows a multinomial distribution with the following probabilities:

$$p_0(t) = \int_0^t \overline{G}_1(t - x)\overline{G}_2(t - x)(dx/t)$$
$$= \int_0^t \overline{G}_1(x)\overline{G}_2(x)(dx/t);$$

$$p_1(t) = \int_0^t \overline{G}_1(t-x)G_2(t-x)(dx/t)$$
$$= \int_0^t \overline{G}_1(x)G_2(x)(dx/t);$$
$$p_2(t) = \int_0^t G_1(t-x)\overline{G}_2(t-x)(dx/t)$$
$$= \int_0^t G_1(x)\overline{G}_2(x)(dx/t).$$

(Note that $1/t$ is the uniform density over $[0, t]$.) Hence, we have,

$$\begin{aligned}
&\mathsf{P}[N_0(t) = n_0, N_1(t) = n_1, N_2(t) = n_2] \\
&= \sum_{n \geqslant n_0+n_1+n_2} \frac{n!}{n_0!n_1!n_2!(n-n_0-n_1-n_2)!} \\
&\quad \times [p_0(t)]^{n_0}[p_1(t)]^{n_1}[p_2(t)]^{n_2} \\
&\quad \cdot [1-p_0(t)-p_1(t)-p_2(t)]^{n-n_0-n_1-n_2} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \\
&= \frac{[\lambda t p_0(t)]^{n_0}[\lambda t p_1(t)]^{n_1}[\lambda t p_2(t)]^{n_2}}{n_0!n_1!n_2!} \\
&\quad \cdot \exp[-\lambda t(p_0(t)+p_1(t)+p_2(t))] \\
&= p(n_0|\lambda t p_0(t)) \cdot p(n_1|\lambda t p_1(t)) \cdot p(n_2|\lambda t p_2(t)). \qquad (8)
\end{aligned}$$

This result indicates that although $X_1(t)$ and $X_2(t)$ are driven by a common arrival process, the three underlying random variables $N_i(t), i = 0, 1, 2$ have independent Poisson distributions with parameters $\lambda t p_i(t)$, $i = 0, 1, 2$, respectively. Thus, $X_1(t)$ and $X_2(t)$ are correlated only because they share a common $N_0(t)$.

Now the joint distribution of $X_1(t)$ and $X_2(t)$ can be expressed through the distributions of $N_i(t), i = 0, 1, 2$. Let $x_1 \wedge x_2 = \min\{x_1, x_2\}$. Then

$$\begin{aligned}
&\mathsf{P}[X_1(t) = x_1, X_2(t) = x_2] \\
&= \sum_{n_0=0}^{x_1 \wedge x_2} \mathsf{P}[N_0(t)+N_1(t) = x_1, N_0(t)+N_2(t) = x_2] \\
&= \sum_{n_0=0}^{x_1 \wedge x_2} \mathsf{P}[N_0(t) = n_0, N_1(t) = x_1-n_0, N_2(t) = x_2-n_0] \\
&= \sum_{n_0=0}^{x_1 \wedge x_2} p(n_0|\lambda t p_0(t))\, p(x_1-n_0|\lambda t p_1(t)) \\
&\quad \times p(x_2-n_0|\lambda t p_2(t)). \qquad (9)
\end{aligned}$$

Thus, due to the special relationship between $X_i(t)$ and $(N_i(t), N_0(t))$, for $i = 1, 2$, all the performance measures of interest can be calculated by first conditioning on $N_0(t)$ and then making use of the independence of $N_1(t)$ and $N_2(t)$.

For the steady-state performance, denote

$$\theta_i := \lim_{t\to\infty} t p_i(t), \qquad i = 0, 1, 2.$$

Letting $t \to \infty$ in (8), and writing $N_i = \lim_{t\to\infty} N_i(t)$ for $i = 0, 1, 2$, we have

$$\begin{aligned}
&\mathsf{P}[N_0 = n_0, N_1 = n_1, N_2 = n_2] \\
&= \frac{(\lambda\theta_0)^{n_0}(\lambda\theta_1)^{n_1}(\lambda\theta_2)^{n_2}}{n_0!n_1!n_2!} \cdot \exp[-\lambda(\theta_0+\theta_1+\theta_2)]. \qquad (10)
\end{aligned}$$

Thus, N_i, $i = 0, 1, 2$, are independent Poisson random variables with parameters $\lambda\theta_i$, $i = 0, 1, 2$, respectively. From (7) this leads to the steady-state limit of $(X_1(t), X_2(t))$, denoted (X_1, X_2):

$$X_1 = N_0 + N_1 = N(\lambda\theta_0) + N(\lambda\theta_1);$$
$$X_2 = N_0 + N_2 = N(\lambda\theta_0) + N(\lambda\theta_2).$$

Thus, X_1 and X_2 are correlated through a common random component $N(\lambda\theta_0)$.

The above analysis extends readily to $m > 2$. In particular, we have

Proposition 1. *For each $i = 1, \ldots, m$, X_i can be expressed as the sum of 2^{m-1} independent Poisson random variables as follows*:

$$X_i = \sum_{S: i\in S} N(\lambda\theta_S),$$

with

$$\theta_S = \int_0^\infty \left[\prod_{k\in S} \overline{G}_k(x)\right]\left[\prod_{j\in\mathcal{F}\setminus S} G_j(x)\right] dx. \qquad (11)$$

Here, $N(\lambda\theta_S)$, $S \subset \mathcal{F}$ is the number of jobs (in steady state) that are still in process with the queues $k \in S$ but have been completed with the queues $j \in \mathcal{F}\setminus S$.

Note that the above proposition holds even when the component lead times are dependent (in which case, use the joint lead time distribution in deriving the probabilities $p_i(t)$); refer to Falin (1994).

For simplicity and without loss of generality, throughout the rest of the paper we shall assume that $\lambda = 1$.

In principle, all the performance measures can be exactly evaluated based on the independent Poisson random variables involved in the joint distribution of $(X_1, \ldots, X_m)$, as per Proposition 1. However, there are $2^m - 1$ such Poisson random variables. Hence, the exponential growth (w.r.t. m) of the number of factors involved in the joint distribution can easily render any exact performance evaluation impractical for systems with a modestly large number of components.

An exception is the special case of deterministic lead times, i.e., $L_i \equiv \ell_i$ for all i. Without loss of generality, assume $\ell_1 < \ell_2 < \cdots < \ell_m$. Then, it can be verified that

$$\begin{aligned}
X_1 &= N(\ell_1) := N_1; \\
X_2 &= N(\ell_1) + N(\ell_2-\ell_1) := N_1 + N_2. \\
&\cdots \\
X_m &= N(\ell_1) + N(\ell_2-\ell_1) + \cdots + N(\ell_m-\ell_{m-1}) \\
&:= N_1 + N_2 + \cdots + N_m, \qquad (12)
\end{aligned}$$

where $N_1, N_2, \ldots, N_m$ are independent Poisson random variables. Thus, in the deterministic lead times case there are only m independent Poisson random variables involved, as opposed to $2^m - 1$ in the case of random lead times.

3. PERFORMANCE ANALYSIS

3.1. Lead Time Variability

First, we investigate the effect on system performance as the lead-time variability decreases. To do so, we compare two systems: the original system with lead time L_i, and another system with lead time $\widetilde{L}_i$, $i = 1, \ldots, m$. Assume $\mathsf{E}[L_i] = \mathsf{E}[\widetilde{L}_i] = \ell_i$, and that L_i is more variable than $\widetilde{L}_i$ in the sense of the "increasing convex ordering," denoted $L_i \geqslant_{icx} \widetilde{L}_i$ (see, e.g., Shaked and Shanthikumar 1994), i.e.,

$$\int_x^\infty \overline{G}_i(u)du \geqslant \int_x^\infty \overline{H}_i(u)du \tag{13}$$

holds for any $x \geqslant 0$. Here, G_i and H_i denote the distribution functions of L_i and $\widetilde{L}_i$, respectively; and $\overline{G}_i = 1 - G_i$, $\overline{H}_i = 1 - H_i$. Note that the above implies, in particular, $\mathsf{Var}[L_i] \geqslant \mathsf{Var}[\widetilde{L}_i]$.

The following inequalities hold (refer to the proof in the Appendix) for the original system and the new system with less variable lead times (denoted with a tilde):

$$\mathsf{P}[X_1 \leqslant x_1, \ldots, X_m \leqslant x_m] \leqslant \mathsf{P}[\widetilde{X}_1 \leqslant x_1, \ldots, \widetilde{X}_m \leqslant x_m], \tag{14}$$

$$\mathsf{P}[X_1 \geqslant x_1, \ldots, X_m \geqslant x_m] \leqslant \mathsf{P}[\widetilde{X}_1 \geqslant x_1, \ldots, \widetilde{X}_m \geqslant x_m], \tag{15}$$

for any $(x_1, \ldots, x_m)$ (vector of nonnegative integers).

The inequality in (14) indicates that $(X_1, \ldots, X_m)$ dominates (i.e., is larger than) $(\widetilde{X}_1, \ldots, \widetilde{X}_m)$ in the *lower orthant* ordering, whereas the inequality in (15) indicates that $(X_1, \ldots, X_m)$ is dominated by (i.e., smaller than) $(\widetilde{X}_1, \ldots, \widetilde{X}_m)$ in the *upper orthant* ordering. Refer to, e.g., Shaked and Shanthikumar (1994).

Let f and $\tilde{f}$ denote the fill rates at the two systems. We have, directly from (14) and (3),

$$f \leqslant \tilde{f}. \tag{16}$$

Let B and $\widetilde{B}$ denote the number of backorders in the two systems in steady state. Then, from the lower orthant ordering in (14), for any positive integer x we have

$$\begin{aligned}\mathsf{P}[B \leqslant x] &= \mathsf{P}[X_1 \leqslant s_1 + x, \ldots, X_m \leqslant s_m + x] \\ &\leqslant \mathsf{P}[\widetilde{X}_1 \leqslant s_1 + x, \ldots, \widetilde{X}_m \leqslant s_m + x] \\ &= \mathsf{P}[\widetilde{B} \leqslant x].\end{aligned}$$

That is,

$$B \geqslant_{\text{st}} \widetilde{B}. \tag{17}$$

To summarize, we have

PROPOSITION 2. *Consider two ATO systems specified above (allowing any number of component queues), one with lead time distributions $G_i, i = 1, \ldots, m$ and the other with less variable lead times with distributions $H_i, i = 1, \ldots, m$. G_i and H_i have the same mean and satisfy* (13). *Everything else is identical. Then, the system with less variable lead times has a higher fill rate, a stochastically smaller number of back orders, and a joint queue-length distribution that is smaller in the lower orthant ordering and larger in the upper orthant ordering.*

3.2. Base-Stock Levels

Next we discuss how the base-stock levels affect performance. From Equations (1) and (2), in particular noticing that $[x]^+$ is a convex function, we can directly establish the following results:

PROPOSITION 3. *B_i is decreasing and convex in s_i, for all i; and B is decreasing in $(s_i)_{i=1}^m$. (Here, decreasing is in the sense of stochastic ordering, e.g., Ross* 1996*, Shaked and Shanthikumar* 1994; *and convex is in the sense of strong stochastic convexity as defined in Shanthikumar and Yao* 1991*a*.) *Consequently, these properties apply to $\mathsf{E}[B_i]$, $i = 1, \ldots, m$, and to $\mathsf{E}[B]$.*

The properties in the above proposition are useful in solving optimal inventory/service trade-off problems. Let c_i be the unit cost of component i. One common formulation is the following (see, e.g., Sherbrooke 1992):

$$\min \mathsf{E}[B] \qquad \text{s.t.} \quad c_1 s_1 + \cdots + c_m s_m \leqslant C, \tag{18}$$

where C is a given positive number representing the limit on the total inventory budget. (Note that in most industrial applications, inventory budget applies to safety stock only. Since the base-stock level can be expressed as the sum of WIP and safety stock, and the WIP part is a constant, the above formulation is consistent with practice.)

Another useful property in solving the optimization problem in (18) is the following (see the Appendix for a proof):

PROPOSITION 4. *Suppose, without loss of generality, that the mean lead times are ordered as follows*:

$$\ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_{m-1} \leqslant \ell_m. \tag{19}$$

If $c_i \geqslant c_j$ for some $i < j$, then the optimal solution to (18) *must satisfy $s_i \leqslant s_j$.*

Despite its deceivingly simple form and the above properties, the problem in (18) is in general intractable because evaluating $\mathsf{E}[B]$ is computationally hard, as discussed in the last section. Below, we develop upper and lower bounds of $\mathsf{E}[B]$ that are easily computable, and use these bounds as surrogates in the optimization problem.

3.3. Bounds

As we have demonstrated in §2, there is a qualitative difference in complexity in evaluating systems with random lead times versus systems with constant lead times: The former involves an exponential number of (independent) Poisson distributions, the latter, a linear number. This motivates

us to use the latter to approximate the former. We shall refer to the system with random lead times as System-r and the system with deterministic lead times as System-d. For each component, the two systems have the same value of mean lead times. To differentiate the two systems, from now on we shall use superscripts r and d for the performance measures of Systems-r and -d, respectively.

Note that System-d corresponds to a system with no lead time variability. In fact, $L_i \geqslant_{icx} \ell_i$. Applying Proposition 2 along with (2), we have the following relations between the backorders in System-r and System-d (where $i = 1, \ldots, m$):

$$\mathsf{E}[B^r] \geqslant \mathsf{E}[B^d] = \mathsf{E}\left[\max_i\{[X_i^d - s_i]^+\}\right] \tag{20}$$

$$\geqslant \max_i\{\mathsf{E}[X_i^d - s_i]^+\}$$

$$= \max_i\{\mathsf{E}[N(\ell_i) - s_i]^+\}, \tag{21}$$

where the second inequality follows from Jensen's inequality (noticing that max is a convex function), and the last equality takes into account that the marginal distributions (for both Systems-r and -d) are equal to $N(\ell_i)$, for $i = 1, \ldots, m$. So, the expected backorders in System-d forms a lower bound for the expected backorders in System-r. Both values also have a common lower bound which is the maximum component-based expected backorders.

To develop an upper bound for $\mathsf{E}[B^r]$, we make use of a simple inequality due first to Lai and Robbins (1976) (to the best of our knowledge). For any set of real values $\{x_1, \ldots, x_m\}$, let $y := \max\{x_1, \ldots, x_m\}$. Then,

$$y \leqslant \alpha + \sum_{i=1}^m (x_i - \alpha)^+, \tag{22}$$

for any real value α.

Because α is arbitrary in (22), we can minimize the right-hand side with respect to α to obtain the best possible bound (among bounds with this form). Therefore, applying this to (2) we have

$$\mathsf{E}[B^r] \leqslant \min_\alpha \left\{\alpha + \sum_{i=1}^m \mathsf{E}[((X_i^r - s_i)^+ - \alpha)^+]\right\}.$$

Let us examine the right-hand side above more closely. Write

$$g(\alpha) := \alpha + \sum_{i=1}^m \mathsf{E}[(\widehat{X}_i - \alpha)^+],$$

where $\widehat{X}_i := [X_i^r - s_i]^+ \geqslant 0$. When $\alpha < 0$, we have

$$g(\alpha) = \alpha + \sum_{i=1}^m [\mathsf{E}(\widehat{X}_i) + |\alpha|] = (m-1)|\alpha| + \sum_{i=1}^m \mathsf{E}(\widehat{X}_i)$$

$$\geqslant \sum_{i=1}^m \mathsf{E}(\widehat{X}_i) = g(0).$$

Hence, the minimal value of $g(\alpha)$ must be obtained at $\alpha \geqslant 0$.

When $\alpha \geqslant 0$, it is easy to verify that

$$\left[(X_i^r - s_i)^+ - \alpha\right]^+ = \left[X_i^r - s_i - \alpha\right]^+.$$

Thus, we have a slightly simplified upper bound:

$$\mathsf{E}[B^r] \leqslant \min_{\alpha \geqslant 0} \left\{\alpha + \sum_{i=1}^m \mathsf{E}\left[(X_i^r - s_i - \alpha)^+\right]\right\}. \tag{23}$$

Proposition 5. *For the expected backorders* $\mathsf{E}[B]$ *in an ATO system with random lead times, we have the lower bounds in* (20) *and* (21) *and the upper bound in* (23).

4. BACK-ORDER MINIMIZATION WITH INVENTORY CONSTRAINT

This section focuses on solving the Optimization Problem (18). Because the objective function $\mathsf{E}[B^r]$ is difficult to evaluate, we replace it by bounds developed in the last section.

4.1. The Lower-Bound Problem

We first consider the problem in (18), with the objective function, $\mathsf{E}[B^r]$, replaced by its lower bound in (21). Denote

$$b_i(s_i) := \mathsf{E}[(N(\ell_i) - s_i)^+] = \ell_i - \sum_{n=0}^{s_i - 1} \overline{P}(n|\ell_i). \tag{24}$$

The resulting optimization problem is:

$$\min \max_i\{b_i(s_i)\} \qquad \text{s.t.} \quad c_1 s_1 + \cdots + c_m s_m \leqslant C. \tag{25}$$

This problem can be solved by a greedy algorithm, which at each step adds one unit to the base-stock level for the component i that has the largest average backorders $b_i(s_i)$, as long as the budget allows. Since $b_i(s_i)$ is decreasing in s_i, this will reduce the largest component under the max operator and hence reduce the objective value. (On the other hand, increasing the base-stock level at any other component has no effect on the objective value.) The algorithm is summarized below:

Algorithm A1

Step 0. For $i = 1, \ldots, m$, set

$$s_i = 0, \qquad b_i = b_i(s_i) = b_i(0) = \ell_i.$$

Set $R = C$.

Step 1. Set

$$b_{\max} = \max_i\{b_i\}, \quad i^* := \arg\max_i\{b_i\}, \quad R \leftarrow R - c_{i^*}.$$

If $R \geqslant 0$, use (24) to update

$$b_{i^*} \leftarrow b_{i^*} - \overline{P}(s_{i^*}|\ell_i), \qquad s_{i^*} \leftarrow s_{i^*} + 1.$$

(Note the s_{i^*} in $\overline{P}(s_{i^*}|\ell_i)$ above is the value *before* the update, $s_{i^*} \leftarrow s_{i^*} + 1$.) Otherwise, stop.

Step 2. Repeat *Step 1*.

In the Appendix we will show:

Proposition 6. *Algorithm A*1 *generates the optimal solution to the problem in* (25).

4.2. The Upper-Bound Problem

With the upper bound in (23) replacing the objective function in (18), and making use of (24), we obtain the following minimization problem:

$$\min\ \alpha+\sum_{i=1}^{m} b_i(s_i+\alpha), \quad \text{s.t.} \quad c_1s_1+\cdots+c_ms_m \leqslant C. \quad (26)$$

Here, α is also a decision variable, in addition to s_i, $i=1,\dots,m$. Because α is part of the argument of $b_i(\cdot)$, we shall treat it as an integer too, just like s_i. Note that we only need to consider $\alpha > 0$, following the analysis in §3.3.

The objective function in (26) is separable and convex in s_i (since $b_i(\cdot)$ is a convex function). It is also convex in α. For any given α, the problem is a bin-packing problem. We propose to use a marginal allocation algorithm to generate a heuristic solution $\mathbf{s}^*(\alpha)$, which at each step gives one unit of C to the index i that incurs the largest decrease in the objective function proportional to the amount of capacity used. Write

$$u(\alpha,\mathbf{s}) := \alpha+\sum_{i=1}^{m} b_i(s_i+\alpha),$$

and

$$u(\alpha)=u(\alpha,\mathbf{s}^*(\alpha)). \quad (27)$$

Note that $\mathbf{s}^*(\alpha)$ is the optimal solution if c_i are the same across i, i.e., $\mathbf{s}^*(\alpha)=\arg\min_{\mathbf{s}} u(\alpha,\mathbf{s})$. In this case, we can show (see the Appendix) that $u(\alpha)$ is convex in α. Thus, we can use a simple line search method to find the optimal α^* that minimizes $u(\alpha)$: Start from $\alpha=0$; increase α one unit at a time; and stop whenever $u(\alpha+1)\geqslant u(\alpha)$. We propose the same procedure as a heuristic for the general case with different c_is.

For any nonnegative integer n, denote

$$\Delta b_i(n) := b_i(n+1)-b_i(n) = -\overline{P}(n|\ell_i), \quad i=1,\dots,m.$$

Because $b_i(\cdot)$ is decreasing convex, $\Delta b_i(\cdot)$ is nonpositive and nondecreasing.

Algorithm A2

Step 0. Set $\alpha=0$, $s_i=0$, $i=1,\dots,m$, $R=C$, $\mathcal{I}=\{1,\dots,m\}$.

Step 1. Identify

$$i^*=\arg\max\left\{-\frac{\Delta b_i(s_i+\alpha)}{c_i};\ i\in\mathcal{I}\right\}$$
$$=\arg\max\{\overline{P}(s_i+\alpha|\ell_i)/c_i;\ i\in\mathcal{I}\},$$

Step 2. If $R\geqslant c_{i^*}$, set $s_{i^*}\leftarrow s_{i^*}+1$. Otherwise, update: $\mathcal{I}\leftarrow\mathcal{I}\setminus\{i^*\}$.

Step 3. Repeat *Steps 1* and *2* until $\mathcal{I}=\varnothing$. Write the objective value as $u(\alpha)$.

Step 4. Set $\alpha\leftarrow\alpha+1$; repeat *Steps 1* through *3* to evaluate $u(\alpha+1)$.

Step 5. Stop if $u(\alpha+1)\geqslant u(\alpha)$, $(\mathbf{s},\alpha)$ being the final solution; else repeat *Step 4*.

PROPOSITION 7. *If c_is are equal for all i, the function $u(\alpha)$ in (27) is convex in α. Consequently, Algorithm A2 solves the problem in (26) optimally in this case.*

4.3. Deterministic Lead Times

Here we focus on the system with deterministic lead times (denoted by a superscript d, as before). We want to solve the following problem:

$$\min\ \mathsf{E}[B^d] \qquad \text{s.t.} \quad c_1s_1+\cdots+c_ms_m\leqslant C, \quad (28)$$

Note that the lower- and upper-bound problems presented in the previous subsections apply here as well. From (20) we know that $\mathsf{E}[B^d]$ is also a lower bound on $\mathsf{E}[B^r]$. In fact, it is a tighter lower bound than (21). So, one may solve (28) and use the optimal solution of this deterministic lead time problem as a heuristic solution of the original random lead time problem. (We consider (21) because it is simpler; only marginal distributions are involved.)

Song (2002) shows that there is a recursive procedure which leads to a closed-form expression for $\mathsf{E}[B^d]$. Specifically, let $\mathbf{x}=(x_1,\dots,x_m)$, $\mathbf{l}=(\ell_1,\dots,\ell_m)$, $\delta_i=\ell_i-\ell_{i-1}$, $i=2,\dots,m$. Let $\mathbf{e}_i$ denote the ith unit vector, $i=1,\dots,m$, and for any j define

$$\mathbf{l}^{(j)}=(\ell_i,\dots,\ell_{j-1},\ell_j,\ell_j,\dots,\ell_j);$$
$$\mathbf{e}^{(j)}=\mathbf{e}_j+\cdots+\mathbf{e}_m;$$
$$x_j^m=\min\{x_j,x_{j+1},\dots,x_m\}.$$

Note that $\mathbf{l}^{(j)}$ is a lead time vector that has equal entries for components j through m. So, $\mathbf{l}^{(m)}=\mathbf{l}$, and $\mathbf{l}^{(1)}$ is a lead time vector with equal entries. Denote by $\mathsf{E}[B^d(\mathbf{x}|\mathbf{l})]$ the expected number of backorders in a system with base-stock level x_i and lead time ℓ_i for item i, $i=1,\dots,m$. When $\ell_1=\ell_2=\cdots=\ell_m=\ell$, we write $\mathsf{E}[B^d(\mathbf{x}|\mathbf{l})]$ as $\mathsf{E}[B^d(\mathbf{x}|\ell)]$. From Song (2002) we have:

$$\mathsf{E}\big[B^d\big(\mathbf{x}|\mathbf{l}^{(j)}\big)\big]=b\big(x_j^m|\delta_j\big)+\ell_j\overline{P}\big(x_j^m-1|\delta_j\big)+\sum_{k=0}^{x_j^m}p(k|\delta_j)\cdot\mathsf{E}\big[B^d\big(\mathbf{x}-k\mathbf{e}^{(j)}|\mathbf{l}^{(j-1)}\big)\big], \quad (29)$$
$$j=m,m-1,\dots 2,$$

$$\mathsf{E}[B^d(\mathbf{x}|\ell)]=\ell-\sum_{k=0}^{\wedge_i x_i-1}\overline{P}(k|\ell). \quad (30)$$

The Recursive Procedure (29) reduces the general problem with different lead times to a problem with equal lead times in $m-1$ steps, at which point the simple Formula (30) applies. For example, in a two-item system with base-stock vector $\mathbf{s}$ and lead time vector $\mathbf{l}$ the procedure yields the following expression:

$$\mathsf{E}[B^d(\mathbf{s}|\mathbf{l})]=b(s_2|\delta_2)+\ell_2\overline{P}(s_2-1|\delta_2)+\sum_{k=0}^{s_2}p(k|\delta_2)\left[\ell_1-\sum_{n=0}^{s_1\wedge(s_2-k)-1}\overline{P}(n|\ell_1)\right].$$

Using the formulas above, one can readily evaluate $\mathsf{E}[B^d]$ for any given $\mathbf{s}$. In the following we present a greedy algorithm, which at each step gives one unit of C to the queue i that achieves the maximal proportional reduction in $\mathsf{E}[B^d]$ value. The algorithm is summarized below:

Algorithm A3

Step 0. Set $s_i = 0$, $i = 1, \ldots, m$, $R = C$, $S = \{1, \ldots, m\}$.

Step 1. Identify

$$i^* := \arg\max \left\{ \frac{\mathsf{E}[B^d(\mathbf{s}|\mathbf{l})] - \mathsf{E}[B^d(\mathbf{s}+\mathbf{e}_i|\mathbf{l})]}{c_i}; \; i \in S \right\}.$$

Step 2. If $R \geqslant c_{i^*}$, update:

$$s_{i^*} \leftarrow s_{i^*} + 1, \qquad R \leftarrow R - c_{i^*}.$$

Otherwise, update: $S \leftarrow S \setminus \{i^*\}$.

Step 3. Repeat *Steps 1* and *2* until $S = \varnothing$.

For small-size problems (in terms of m and C), the optimal solution to the problem in (28) can be generated by complete enumeration. In our numerical experiments the solutions generated by the greedy heuristic in Algorithm A3 coincide with complete enumeration in most cases, although exceptions do exist.

5. INVENTORY MINIMIZATION WITH A FILL-RATE CONSTRAINT

In this section we study the problem of minimizing the total inventory cost over all m components, subject to a required fill rate, β (e.g., $\beta = 95\%$). Let h_i be the unit inventory holding cost at queue i; then the problem is

$$\min \sum_{i=1}^{m} h_i \mathsf{E}[I_i] = \sum_{i=1}^{m} h_i \mathsf{E}[(s_i - X_i)^+], \quad \text{s.t. } f^r \geqslant \beta. \tag{31}$$

Clearly, the objective function is increasing and convex, as well as separable, in s_i.

Note that whereas f^r is computationally intractable, it has the following lower bound:

$$f^r \geqslant \mathsf{P}\big[X_1^r < s_1\big] \cdots \mathsf{P}\big[X_m^r < s_m\big].$$

To justify the above inequality, note that for each i, X_i can be expressed as a sum of independent Poisson random variables. Therefore, $(X_1^r, \ldots, X_m^r)$ are *associated* random variables, and hence the above inequality. (Refer to Shaked and Shanthikumar 1994. Also refer to Song (1998) for a similar result.)

Hence, replace f^r by its lower bound in (31), and taking logarithms on both sides of the constraint, we have:

$$\min \sum_{i=1}^{m} h_i \mathsf{E}[(s_i - X_i)^+], \quad \text{s.t. } \sum_{i=1}^{m} -\log P(s_i - 1|\ell_i) \leqslant -\log\beta. \tag{32}$$

Note that $P(s_i|\ell_i)$ is increasing in s_i; and when $s_i \geqslant \ell_i$, it is also concave in s_i. (When $s_i = \ell_i$, $P(s_i|\ell_i)$ is about $1/2$, following a normal approximation. Hence, for moderately high fill rate β, $s_i \geqslant \ell_i$ is the right range to focus on.) Because an increasing and concave function is log-concave, the left-hand side of the constraint in (32) is decreasing and convex, as well as separable, in s_i.

Therefore, (32) is a separable convex programming problem which can be solved via a greedy algorithm as follows. From (5) and (6), we have

$$\mathsf{E}[I_i] = \sum_{n=0}^{s_i - 1} P(n|\ell_i). \tag{33}$$

Thus, the problem can be rewritten as

$$\min \sum_{i=1}^{m} h_i \left[\sum_{n=0}^{s_i - 1} P(n|\ell_i) \right], \quad \text{s.t. } \sum_{i=1}^{m} -\log P(s_i - 1|\ell_i) \leqslant -\log\beta. \tag{34}$$

Note that increasing s_i to $s_i + 1$ increases the objective function by: $h_i P(s_i|\ell_i)$; whereas it decreases the left-hand side of the constraint by $\log P(s_i|\ell_i) - \log P(s_i - 1|\ell_i)$. The greedy algorithm below favors—at each step when s_i, for some i, is increased—a small increase in the objective value and a large decrease in the left-hand side of the constraint.

Algorithm A4

- For $i = 1, \ldots, m$, set $s_i = \ell_i$.
- Identify

$$i^* = \arg\min \left\{ \frac{h_i P(s_i|\ell_i)}{\log P(s_i|\ell_i) - \log P(s_i - 1|\ell_i)} \right\};$$

and set $s_{i^*} \leftarrow s_{i^*} + 1$.

- Repeat the above until the constraint in (34) is satisfied.

Now consider the special case of deterministic lead times. Because the objective function in (34) is separable (i.e., only the marginal distributions are used) it remains unchanged when the lead times are deterministic. As to the constraint, an exact evaluation of f^d is tractable. In fact, assuming $\ell_1 < \ell_2 < \cdots < \ell_m$ and using (12), we have

$$\begin{aligned} f^d &= \mathsf{P}\big[X_1^d < s_1, \ldots, X_m^d < s_m\big] \\ &= \mathsf{P}[N_1 < s_1, N_1 + N_2 < s_2, \ldots, N_1 + \cdots + N_m < s_m] \\ &= \sum_{n_1 < \min_k s_k} p(n_1|\ell_1) \sum_{n_2 < \tilde{s}_2} p(n_2|\delta_2) \cdots \\ &\quad \times \sum_{n_{m-1} < \tilde{s}_{m-1}} p(n_{m-1}|\delta_{m-1}) P(\tilde{s}_m|\delta_m), \end{aligned} \tag{35}$$

where for $i \geqslant 2$,

$$\delta_i = \ell_i - \ell_{i-1}, \qquad \tilde{s}_i = \min_{k \geqslant i} s_k - n_1 - \cdots - n_{i-1}.$$

In principle, we can evaluate the increase of the above when s_i, for some i, is increased by one unit, and divide this incremental increase by the incremental increase in the

Table 1. Solution comparison: Lower and upper bound algorithms for minimizing backorders. $\mathsf{E}[L] = (1, 2, 3, 4)$, $\lambda = 2$, $c = (1, 1, 1, 1)$.

	Constraint	Solution				Average Backorders $\mathsf{E}[B]$ Lead Time Distribution			
Algorithm	C	$s1$	$s2$	$s3$	$s4$	Deterministic	Uniform	Erlang-2	Exponential
Lower Bound (A1)	20	2	4	6	8	1.5325	1.5869	1.7688	1.8921
Upper Bound (A2)	20	2	4	6	8	1.5325	1.5869	1.7688	1.8921
Lower Bound (A1)	25	2	5	8	10	0.8175	0.9073	0.9945	1.0616
Upper Bound (A2)	25	3	5	7	10	0.8069	0.8374	0.9589	1.0348
Lower Bound (A1)	30	3	6	9	12	0.4019	0.4137	0.4629	0.4975
Upper Bound (A2)	30	4	6	9	11	0.3755	0.3928	0.4485	0.4855
Lower Bound (A1)	35	4	8	10	13	0.1602	0.1670	0.1853	0.1983
Upper Bound (A2)	35	5	7	10	13	0.1508	0.1582	0.1823	0.1980

Note: Lower Bound: Greedy method for $\min\{\max(\mathsf{E}[B_i])\}$; Algorithm A1.
Upper Bound: Greedy method for $\min\{u(a, s)\}$; Algorithm A2.

objective function, $h_i P(s_i|\ell_i)$, and use this ratio to identify i^*, which corresponds to the largest such ratio. The downside is that computing the incremental increase of f^d from (35) is significantly harder than evaluating $\log P(s_i|\ell_i) - \log P(s_i - 1|\ell_i)$. Furthermore, although we deal with the exact constraint, the optimality of the greedy scheme is not guaranteed, as (35) is nonseparable, and it is difficult to tell what properties it possesses with respect to $(s_i)_{i=1}^m$. Therefore, even in the case of deterministic lead times there are still advantages in using the lower-bound constraint.

6. NUMERICAL EXAMPLES

We have conducted numerical experiments to test the effectiveness of the bounds and the surrogate optimization problems. The numerical examples focus on a four-component system where customer orders follow a Poisson process with rate $\lambda = 2$. The lead times for each component i are i.i.d. random variables with a mean $\mathsf{E}[L_i] = \ell_i = i$, for $i = 1, 2, 3, 4$. The inventory of component i is controlled by a base-stock policy with base-stock level s_i. Four types of lead time distribution are considered:

1. Deterministic: $L_i = \ell_i$.
2. Uniform$(\ell_i/2,\ 3\ell_i/2)$ with mean ℓ_i and variance $\ell_i^2/12$.
3. Erlang$(2/\ell_i,\ 2)$ with mean ℓ_i and variance $\ell_i^2/4$.
4. Exponential$(1/\ell_i)$ with mean ℓ_i and variance ℓ_i^2.

Thus, the mean lead times are the same in all cases, but the variances are different.

Table 1 compares the lower- and upper-bound approaches to the backorder minimization problem. We focus on the case in which $c_i = 1$ for all i. Recall that in this setting both greedy Algorithms A1 and A2 generate the optimal solution for the corresponding surrogate problems. Also recall that both bounds use the marginal distributions of the outstanding orders X_i, $i = 1, 2, 3, 4$. Therefore, the bounds apply to any lead times, random or deterministic. In Table 1, we first report the solutions (base-stock levels) generated by the two approaches for four values of the inventory budget constraint C. We then report the simulated expected backorders $\mathsf{E}[B]$ using these base-stock policies. That is, we evaluate the objective function of the original problem at the solutions generated by the surrogate problems. The simulation results are based on 1,000 regenerative cycles. The confidence intervals are within the 4th decimal place. The highlighted numbers indicate the smaller value of the optimal objective values. From these examples we observe that the upper-bound approach always dominates the lower-bound approach. This observation remains true for our other experiments with equal c_i, of which Table 1 is a sample. Also, we do observe that the value of $\mathsf{E}[B]$ increases as the lead time becomes more variable across the systems, as stated in Proposition 2.

Table 2 compares the results from solving Problem (28) (i.e., System-d) with those from using the lower- and upper-bound approaches. Again, we set $c_i = 1$ for all i. For System-d we use Algorithm A3, the greedy method to solve Problem (28). Similar to Table 1, we evaluate each approach through simulating the $\mathsf{E}[B]$ value corresponding to the solution generated. We also compare these against the optimal solution of the original problem generated by complete enumeration. The highlighted numbers indicate the optimal objective values for each system; the corresponding solutions are the respective optimal policies. In most cases, the optimal solution coincides with those generated by Algorithm A2 (the upper-bound approach) or Algorithm A3 (the System-d approach), whichever gives a lower $\mathsf{E}[B^r]$. For example, in Table 2, $C = 40$, the complete enumeration shows that the solution generated by Algorithm A3 is optimal for systems with deterministic lead time, while the solution generated by Algorithm A2 is optimal for systems with other lead time distributions. However, exceptions do exist, as in Table 2, $C = 15$. Here, Algorithm A3 generates a solution $(1, 3, 4, 7)$. While the complete enumeration confirms that this is in fact optimal for the deterministic-lead time system, a different solution, $(1, 2, 5, 7)$, turns out to be optimal for systems with other

Table 2. Solution comparison: Algorithms for minimizing backorders. $\mathsf{E}[L] = (1, 2, 3, 4)$, $\lambda = 2$, $c = (1, 1, 1, 1)$.

	Constraint	Solution				Average Backorders $\mathsf{E}[B]$ Lead Time Distribution			
Algorithm	C	$s1$	$s2$	$s3$	$s4$	Deterministic	Uniform	Erlang-2	Exponential
Lower Bound (A1)	15	0	3	5	7	2.7198	2.7645	2.9293	3.0541
Upper Bound (A2)	15	0	3	5	7	2.7198	2.7645	2.9293	3.0541
Greedy-d (A3)	15	1	3	4	7	2.6152	2.6867	2.9217	3.0756
Enumeration	15	1	2	5	7	2.6152*	2.6633	2.8943	3.0470
Lower Bound (A1)	30	3	6	9	12	0.4019	0.4136	0.4618	0.4975
Upper Bound (A2)	30	4	6	9	11	0.3775	0.3936	0.4493	0.4855
Greedy-d (A3)	30	4	6	9	11	0.3775	0.3936	0.4493	0.4855
Lower Bound (A1)	40	5	9	12	14	0.0568	0.0588	0.0650	0.0694
Upper Bound (A2)	40	5	9	12	14	0.0568	0.0588	0.0650	0.0694
Greedy-d (A3)	40	6	9	11	14	0.0554	0.0591	0.0677	0.0727

Note: Greedy-d: Greedy method for $\min(\mathsf{E}[B^d])$; Algorithm A3.
Lower Bound: Greedy method for $\min\{\max(\mathsf{E}[B_i])\}$; Algorithm A1.
Upper Bound: Greedy method for $\min\{u(a, s)\}$; Algorithm A2.
Optimal: Highlighted numbers; complete enumeration.
* This value corresponds to the solution (1,3,4,7), generated by enumeration, as well as A3.

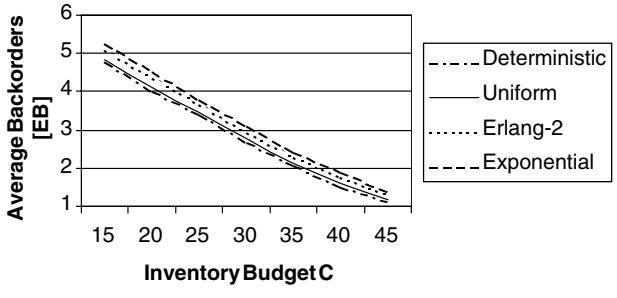
lead time distributions. Indeed, evaluating $(1, 2, 5, 7)$ in the deterministic-lead time system we see a slightly higher objective value, 2.6193, than the optimal value 2.6152 of solution $(1, 3, 4, 7)$.

Table 3 repeats the experiment in Table 2 for a case of different component costs: $c_1 = 1, c_2 = 2, c_3 = 1, c_4 = 3$. When the c_is are different, Algorithm A3 appears to be the most effective. Indeed, in almost all examples we have tried, the optimal solutions obtained through complete enumeration (highlighted in the table) coincide with those generated from Algorithm A3. Unlike the case of equal c_is, the lower-bound approach (A1) appears to be more effective than the upper-bound approach (A2) here. This is perhaps due to the fact that in this case Algorithm A2 is no longer guaranteed to be optimal for the upper-bound problem.

It is worth mentioning that it takes only a few seconds to generate *all* the greedy solutions reported here, but it takes as long as 30 minutes to an hour to just *evaluate* the objective value of *each* solution via simulation.

Figure 1 summarizes the optimal inventory and service trade-off in systems with different lead times, using the data

Figure 1. Optimal inventory-service trade-off: Backorders.



in Table 3. Here, "inventory" is represented by the component inventory budget C, and "service" is represented by the minimum average product backorders that can be achieved within budget limit C. (Note that according to Little's law, the average number of backorders is proportional to the average delay before the order can be filled.) It is interesting to observe the following: (a) the shape of the optimal trade-off curve is rather insensitive to lead time distributions; and (b) the curve is nearly linear, which implies that the average backorders (or equivalently, the average customer waiting time) improves at a nearly constant rate as the inventory budget increases. Furthermore, to determine the rate of improvement, based on (a), we can use System-d (deterministic lead times), for which the solutions are easily generated by the greedy Algorithm A3.

Table 4 illustrates the effectiveness of Algorithm A4 for the optimization problem with service constraint. It presents the solutions and the objective values of the surrogate problem corresponding to different values of the service level β. It also reports the corresponding order fill rates and component fill rates in the original system. The order fill rates for systems with random lead times are evaluated by simulation. The other values are from exact computation. As the numbers show, the order fill rates are much lower than the component fill rates. Recall that the lower bound of order fill rate is the product of the component fill rates, which is used in Algorithm A4. Note that this lower-bound approach is rather conservative, in the sense that its solution usually results in an order fill rate (in the original system) that is substantially higher than the required service level. Therefore, the solutions of Algorithm A4 are likely to be suboptimal and Table 4 provides some examples in this regard. Here, we focus on Erlang-2 lead times. For each service level, the table shows the best solution (obtained from a search around the solution generated by

Table 3. Solution comparison: Algorithms for minimizing backorders. $\mathsf{E}[L] = (1, 2, 3, 4)$, $\lambda = 2$, $c = (1, 2, 1, 3)$.

	Constraint	Solution				Average Backorders $\mathsf{E}[B]$ Lead Time Distribution			
Algorithm	C	$s1$	$s2$	$s3$	$s4$	Deterministic	Uniform	Erlang-2	Exponential
Greedy-A1	15	0	0	2	4	4.9321	5.0189	5.3224	5.5006
Greedy-A2	15	0	1	4	3	5.1421	5.1553	5.2941	5.4147
Greedy-A3	15	0	0	3	4	4.8008	4.8515	5.0956	5.2590
Greedy-A1	20	0	1	3	5	4.0056	4.0908	4.3753	4.5525
Greedy-A2	20	0	1	6	4	4.4067	4.4192	4.5565	4.6735
Greedy-A3	20	0	1	3	5	4.0056	4.0908	4.3753	4.5525
Greedy-A1	25	0	1	4	6	3.5761	3.6266	3.8654	4.0284
Greedy-A2	25	0	2	6	5	3.6091	3.6168	3.7517	3.8671
Greedy-A3	25	1	2	5	5	3.4341	3.4604	3.6549	3.8000
Greedy-A1	30	0	2	5	7	2.8955	2.9362	3.1286	3.2709
Greedy-A2	30	1	2	7	6	2.9040	2.9229	3.0672	3.1866
Greedy-A3	30	1	3	5	6	2.6931	2.7486	2.9608	3.1070
Greedy-A1	35	1	3	5	7	2.3165	2.3823	2.5992	2.7486
Greedy-A2	35	2	4	7	6	2.4327	2.4398	2.5310	2.6184
Greedy-A3	35	2	3	6	7	2.0623	2.1037	2.2941	2.4294
Greedy-A1	40	2	4	6	8	1.5325	1.5845	1.7694	1.8900
Greedy-A2	40	3	4	8	7	1.7822	1.7922	1.8858	1.9656
Greedy-A3	40	2	4	6	8	1.5325	1.5845	1.7694	1.8900
Greedy-A1	45	2	4	7	9	1.2276	1.2599	1.4005	1.5002
Greedy-A2	45	3	5	8	8	1.2109	1.2275	1.3143	1.3835
Greedy-A3	45	3	4	7	9	1.1243	1.1600	1.3001	1.3934

Note: Greedy-A1: Greedy method for the lower bound min{u(a,s)}; Algorithm A2.
Greedy-A2: Greedy method for the lower bound problem min{max(E[Bi])}; Algorithm A1.
Greedy-A3: Greedy method for min(E[B]) in the deterministic lead time system; Algorithm A3.
Optimal: The highlighted numbers; complete enumeration.

Algorithm A4). For reference, the table also reports the lower bound, the item fill rates, and the order fill rates under other lead time distributions, using the best solution under Erlang-2 lead times.

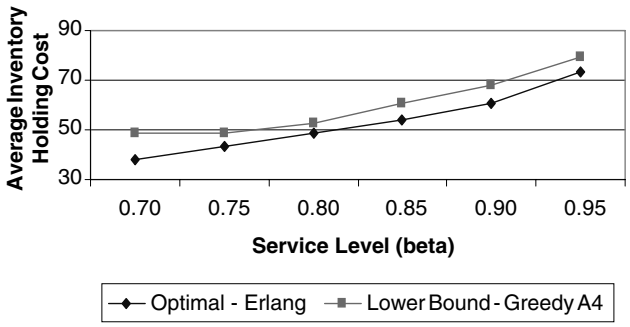
Note, however, that Algorithm A4 has considerable advantage in computation time. The solutions reported here only take a few seconds to generate. Therefore, it can be used to quickly generate an initial solution, followed by a

Table 4. Numerical examples for Algorithm 4: Minimize average inventory cost subject to a fill rate constraint. $\mathsf{E}[L] = (1, 2, 3, 4)$, $\lambda = 2$, $h = (1, 3, 3, 5)$.

Service Level	Solution				Objective	Order Fill Rate					Component Fill Rate			
β	$s1$	$s2$	$s3$	$s4$	Value	LowerBd	Exponential	Erlang	Uniform	Deterministic	$F1$	$F2$	$F3$	$F4$
0.70	6	8	10	12	48.9879	0.7592	0.8104	0.8244	0.8482	0.8549	0.9834	0.9489	0.9161	0.8881
	5	7	9	11	37.9693	0.5824	0.6841	0.7050	0.7408	0.7520	0.9473	0.8893	0.8472	0.8159
0.75	6	8	10	12	48.9879	0.7592	0.8104	0.8244	0.8482	0.8549	0.9834	0.9489	0.9161	0.8881
	6	7	9	12	43.3931	0.6581	0.7343	0.7542	0.7863	0.7958	0.9834	0.8893	0.8472	0.8881
0.80	7	8	11	12	52.8555	0.8031	0.8388	0.8495	0.8652	0.8696	0.9955	0.9489	0.9574	0.8881
	6	8	10	12	48.9879	0.7592	0.8103	0.8243	0.8472	0.8549	0.9834	0.9489	0.9161	0.8881
0.85	7	9	11	13	60.4725	0.8732	0.8956	0.9028	0.9155	0.9202	0.9955	0.9786	0.9574	0.9362
	6	8	10	13	53.6690	0.8003	0.8436	0.8555	0.8752	0.8817	0.9834	0.9489	0.9161	0.9362
0.90	7	9	12	14	68.2413	0.9220	0.9354	0.9403	0.9477	0.9504	0.9955	0.9786	0.9799	0.9658
	7	9	11	13	60.4725	0.8732	0.8956	0.9031	0.9154	0.9202	0.9955	0.9786	0.9574	0.9362
0.95	7	10	13	15	79.1041	0.9618	0.9674	0.9697	0.9734	0.9746	0.9955	0.9919	0.9912	0.9827
	7	10	12	15	73.1550	0.9508	0.9505	0.9525	0.9667	0.9686	0.9955	0.9919	0.9799	0.9827

Note: The shaded area reports the optimal solution for the system with Erlang-2 lead times.

Figure 2. Optimal inventory-service trade-off: Fill rate.



neighborhood search to find the best solution, as done in Table 4.

Figure 2 shows the minimum inventory holding costs corresponding to several service-level (fill rate) requirements, using the optimal solution for the Erlang-2 case reported in Table 4. Observe that as the service level increases, the optimal inventory cost increases at an increasing rate. For each service-level requirement, we also plot the corresponding objective value generated by Algorithm A4. Obviously, this curve is a lower bound of the true optimal trade-off curve. However, it follows quite closely the rate of change in the optimal curve.

7. CONCLUDING REMARKS

In this study we have carried out an exact performance analysis of a single-product ATO system under base-stock control and with i.i.d. component lead times. The results shed light on how system parameters affect performance, and also lead to performance bounds that are easy to compute. Several optimization models are developed based on these performance bounds. Greedy-type algorithms are shown to be effective in generating solutions to the optimal inventory-service trade off. The extension to systems with multiple products turns out to be far from routine. New approaches are needed for both performance analysis and optimization. These will be the focus of our follow-up studies.

APPENDIX

PROOF OF PROPOSITION 2. From the discussions preceding Proposition 2, we know it suffices to show the inequality in (14). For ease of exposition, we focus on the case of $m = 2$; the argument below extends readily to the general case. We shall use the same notation as in §2, and use "tilde" to denote the new system with reduced lead time variability.

First note that the inequality in (13) becomes an equality when $x = 0$ as both sides are equal to ℓ_i. So, we have, for any $x \geqslant 0$,

$$\int_0^x \overline{G}_i(u)du \leqslant \int_0^x \overline{H}_i(u)du \quad i = 1, 2. \tag{36}$$

Making use of the above, along with integration by parts, we have

$$\begin{aligned}
\theta_0 &= \int_0^\infty \overline{G}_1(x)\overline{G}_2(x)dx = \int_0^\infty \overline{G}_1(x)d\left[\int_0^x \overline{G}_2(u)du\right] \\
&= \int_0^\infty \left[\int_0^x \overline{G}_2(u)du\right] dG_1(x) \\
&\leqslant \int_0^\infty \left[\int_0^x \overline{H}_2(u)du\right] dG_1(x) = \int_0^\infty \overline{G}_1(x)\overline{H}_2(x)dx \\
&\leqslant \int_0^\infty \overline{H}_1(x)\overline{H}_2(x)dx = \tilde{\theta}_0.
\end{aligned}$$

Here, inequality follows from (36), since $dG_1(x) \geqslant 0$, and the second inequality is similarly established as the first one.

Hence, letting $\epsilon = \tilde{\theta}_0 - \theta_0 \geqslant 0$, and taking into account

$$\theta_0 + \theta_i = \tilde{\theta}_0 + \tilde{\theta}_i = \ell_i, \qquad i = 1, 2,$$

we can write

$$\theta_1 = \tilde{\theta}_1 + \epsilon, \qquad \theta_2 = \tilde{\theta}_2 + \epsilon.$$

This leads to

$$\begin{aligned}
X_1 &= N(\theta_0) + N(\tilde{\theta}_1) + N(\epsilon), \\
X_2 &= N(\theta_0) + N(\tilde{\theta}_2) + \widehat{N}(\epsilon).
\end{aligned}$$

All Poisson random variables involved in each of the summations above are independent of one another; $\widehat{N}(\epsilon)$ denotes an *independent* replica of $N(\epsilon)$, whereas without the "hat", $N(\theta_0)$ for instance, which appears in both X_1 and X_2, denotes the *same* random variable. On the other hand, for the new system, we have

$$\begin{aligned}
\widetilde{X}_1 &= N(\tilde{\theta}_0) + N(\tilde{\theta}_1) = N(\theta_0) + N(\epsilon) + N(\tilde{\theta}_1), \\
\widetilde{X}_2 &= N(\tilde{\theta}_0) + N(\tilde{\theta}_2) = N(\theta_0) + N(\epsilon) + N(\tilde{\theta}_2).
\end{aligned}$$

Comparing (X_1, X_2) and $(\widetilde{X}_1, \widetilde{X}_2)$, we note that the only difference is that the $N(\epsilon)$ term is the common random variable in both $\widetilde{X}_1$ and $\widetilde{X}_2$, whereas in the original system, it becomes two independent replicas. Now, since

$$\begin{aligned}
\mathsf{P}[N(\epsilon) \leqslant y_1, N(\epsilon) \leqslant y_2] &= \mathsf{P}[N(\epsilon) \leqslant y_1 \wedge y_2] \\
&= \mathsf{P}[N(\epsilon) \leqslant y_1] \wedge \mathsf{P}[N(\epsilon) \leqslant y_2] \\
&\geqslant \mathsf{P}[N(\epsilon) \leqslant y_1] \cdot \mathsf{P}[N(\epsilon) \leqslant y_2] \\
&= \mathsf{P}[N(\epsilon) \leqslant y_1, \widehat{N}(\epsilon) \leqslant y_2],
\end{aligned}$$

conditioning on the other independent Poisson random variables (i.e., other than those associated with ϵ), we have:

$$\mathsf{P}[X_1 \leqslant x_1, X_2 \leqslant x_2] \leqslant \mathsf{P}[\widetilde{X}_1 \leqslant x_1, \widetilde{X}_2 \leqslant x_2],$$

for any x_1 and x_2 (nonnegative integers).

A similar argument may be used to show (15). In both cases, the proof hinges on the fact that in the tilde system probability is moved from the individual component to the

joint queues, implying that the items of each type that are waiting are more positively correlated. □

PROOF OF PROPOSITION 4. Consider again a two-component system. Since $c_1 \geqslant c_2$, if (s_2, s_1) with $s_1 < s_2$ is a feasible solution, then (s_1, s_2) is also a feasible solution. Thus, it is sufficient to show

$$\mathsf{E}[B(s_1, s_2)] \leqslant \mathsf{E}[B(s_2, s_1)], \tag{37}$$

for $s_1 < s_2$.

Consider two generic discrete random variables, X_1 and X_2, which are not necessarily independent. Following Shanthikumar and Yao (1991b), X_1 and X_2 satisfy the (joint) likelihood ratio ordering, denoted $X_1 \leqslant_{\text{lr:j}} X_2$, if

$$\mathsf{P}[X_1 = x_2, X_2 = x_1] \leqslant \mathsf{P}[X_1 = x_1, X_2 = x_2],$$

for any $x_1 \leqslant x_2$. In Shanthikumar and Yao (1991b), it is shown that this ordering is preserved by the class of bivariate functions $g(x_1, x_2)$ that satisfy

$$g(x_1, x_2) \leqslant g(x_2, x_1), \quad \forall\, x_1 \leqslant x_2. \tag{38}$$

That is,

$$X_1 \leqslant_{\text{lr:j}} X_2 \qquad \text{iff} \qquad \mathsf{E}g(X_1, X_2) \leqslant \mathsf{E}g(X_2, X_1),$$

for any $g(x_1, x_2)$ that satisfies (38).

From the analysis in §2, we can write

$$X_1 = N(\theta_0) + N(\theta_1) := N_0 + N_1,$$
$$X_2 = N(\theta_0) + \widetilde{N}(\theta_1) + N(\delta) := N_0 + \widetilde{N}_1 + N_2,$$

where $\delta = \ell_2 - \ell_1 = \theta_2 - \theta_1$. Recall that $\widetilde{N}_1$ is an independent replica of N_1; and $\{N_0, N_1, \widetilde{N}_1, N_2\}$ are all independent.

Consider $x_1 \leqslant x_2$. We have

$$\begin{aligned}
&\mathsf{P}[X_1 = x_2, X_2 = x_1 \,|\, N_0 = n_0, N_2 = n_2] \\
&\quad = \mathsf{P}[N_1 = x_2 - n_0]\mathsf{P}[\widetilde{N}_1 = x_1 - n_0 - n_2] \\
&\quad \leqslant \mathsf{P}[N_1 = x_1 - n_0]\mathsf{P}[\widetilde{N}_1 = x_2 - n_0 - n_2] \\
&\quad = \mathsf{P}[X_1 = x_1, X_2 = x_2 \,|\, N_0 = n_0, N_2 = n_2],
\end{aligned}$$

where the inequality can be directly verified from the Poisson distribution. In fact, after canceling out common terms on both sides, the inequality is reduced to the following:

$$(x_2 - n_0)!(x_1 - n_0 - n_2)! \geqslant (x_1 - n_0)!(x_2 - n_0 - n_2)!,$$

which obviously holds, following $x_1 \leqslant x_2$. Hence, unconditioning, we get

$$\mathsf{P}[X_1 = x_2, X_2 = x_1] \leqslant \mathsf{P}[X_1 = x_1, X_2 = x_2].$$

That is, $X_1 \leqslant_{\text{lr:j}} X_2$.

Now, similar to the equation in (20), we can write

$$\mathsf{E}[B(s_1, s_2)] = \mathsf{E}\max\{[X_1 - s_1]^+, [X_2 - s_2]^+\}.$$

Consider $s_1 \leqslant s_2$, and the bivariate function

$$g(x_1, x_2) := \max\{[x_1 - s_1]^+, [x_2 - s_2]^+\}.$$

Then, direct verification shows that (38) does hold in this case. Hence, $X_1 \leqslant_{\text{lr:j}} X_2$ implies $\mathsf{E}g(X_1, X_2) \leqslant \mathsf{E}g(X_2, X_1)$, which is exactly what is desired in (37).

In fact, the g function above belongs to the class of functions that preserves the *arrangement ordering* (Shaked and Shanthikumar 1994, Shanthikumar and Yao 1991b), which is essentially a pairwise ordering. Hence, no generality is lost in our argument above focusing on the case of $m = 2$. More specifically, when there are $m > 2$ components, we only need to focus on two of them, say, X_1 and X_2; and the above argument will go through by conditioning on all the other Poisson variables not involved in X_1 and X_2. □

PROOF OF PROPOSITION 6. Suppose $\mathbf{s}^* = (s_1^*, \ldots, s_m^*)$ is an optimal solution to the problem in (25) (recall that the optimal solution always exists). Let $\mathbf{s} = (s_1, \ldots, s_m)$ be the solution obtained through Algorithm A1. We want to show that $\mathbf{s}$ and $\mathbf{s}^*$ must have the same objective value.

First, if $s_i \geqslant s_i^*$ for all i, then by the decreasing property of $b_i(\cdot)$ we have $b_i(s_i) \leqslant b_i(s_i^*)$. Therefore $\max_i\{b_i(s_i)\} \leqslant \max_i\{b_i(s_i^*)\}$. From the optimality of $\mathbf{s}^*$, we also have the reversed inequality. Thus, the two values must be equal, that is, $\mathbf{s}$ and $\mathbf{s}^*$ have the same objective value, implying that $\mathbf{s}$ is optimal.

Now suppose that there exists some i such that $s_i < s_i^*$. For simplicity, suppose $i = 1$. Then, we need to consider two possibilities:

Case 1. There exists some $j > 1$ such that $s_j > s_j^*$. For simplicity, suppose $j = 2$. Recall that Algorithm A1 increases the base-stock levels one unit at a time. Assume we reached s_2 from $s_2 - 1$ in step n for some n. Then at step $n-1$ the solution is of the form $(s_1', s_2 - 1, s_3', \ldots, s_m')$ where $s_i' \leqslant s_i$ for $i \neq 2$. By the construction of the algorithm, we must have $b_2(s_2 - 1) = \max\{b_2(s_2 - 1), b_i(s_i'), i \neq 2\}$. Moreover, due to the decreasing property of $b_i(\cdot)$, we have

$$b_2(s_2 - 1) > b_2(s_2) \quad \text{and} \quad b_i(s_i') \geqslant b_i(s_i), \quad i \neq 2.$$

So,

$$\begin{aligned}
b_2(s_2 - 1) &= \max\{b_2(s_2 - 1), b_i(s_i'), i \neq 2\} \\
&> \max\{b_2(s_2), b_i(s_i'), i \neq 2\} \geqslant \max_i\{b_i(s_i)\}.
\end{aligned}$$

Note that $s_2 - 1 \leqslant s_2^*$ implies $b_2(s_2^*) \geqslant b_2(s_2 - 1)$, which leads to

$$\max_i\{b_i(s_i^*)\} \geqslant b_2(s_2^*) \geqslant b_2(s_2 - 1) > \max_i\{b_i(s_i)\},$$

contradicting the optimality of $\mathbf{s}^*$. Thus, we must have $s_j \leqslant s_j^*$ for all $j > 1$, which is Case 2 below.

Case 2. $s_1 < s_1^*$ and $s_j \leqslant s_j^*$ for all $j > 1$. In this case, $(s_1 + 1, s_2, \ldots, s_m)$ is also a feasible solution. By the principles of Algorithm A1, $b_1(s_1) < \max_i\{b_i(s_i)\}$, because otherwise the algorithm would lead to solution

$(s_1+1, s_2, \ldots, s_m)$. This implies that $(s_1+1, s_2, \ldots, s_m)$ gives the same objective value as $\mathbf{s}$. We can keep increasing the value of s_1 to s_1^* in the same fashion, and applying the same logic, conclude that $(s_1^*, s_2, \ldots, s_m)$ has the same objective value as $\mathbf{s}$. Using the same argument for the other components where $s_j < s_j^*$, we can show that $\mathbf{s}^*$ has the same objective value as $\mathbf{s}$, therefore $\mathbf{s}$ is also optimal. This completes the proof. □

Proof of Proposition 7. Since we only consider the case with equal c_is here, without loss of generality we assume $c_i = 1$ for all i. Due to the monotonicity of $b_i(\cdot)$, the budget constraint must be tight at optimality. Choose any α and C, and let $\mathbf{s}_C(\alpha) = (s_1, \ldots, s_m)$ be the optimal allocation corresponding to α and C. Accordingly, we can write

$$u_C(\alpha) := u(\alpha) = \alpha + \sum_i b_i(s_i + \alpha).$$

Also, $s_1 + \cdots + s_m = C$. Now consider $\alpha + 1$. Note that we can rewrite the right-hand side above as follows:

$$\alpha + 1 + \sum_i b_i(s_i - 1 + \alpha + 1) - 1.$$

We show below that the allocation $(s_1 - 1, \ldots, s_m - 1)$ is optimal with respect to $\alpha + 1$ and the first $C' := C - m$ units of the total budget C. Clearly, this is a feasible solution. (Note that if $s_i = 0$, then $s_i - 1 = -1$, and $b_i(-1) = \ell_i + 1$, following the first equality in (24). It is also worth noting that in principle the greedy Algorithm A2 should start from sufficiently negative s_i values. The choice of a zero initial solution in A2 is based on practicality, as negative base-stock levels are not likely to be candidates for optimality in the kind of applications we are concerned with.)

To argue that $(s_1 - 1, \ldots, s_m - 1)$ is optimal, use contradiction. If it is not optimal, suppose $(s_1' - 1, \ldots, s_m' - 1)$ is optimal instead. Then, we must have

$$\begin{aligned} u_{C'}(\alpha + 1) &= \alpha + 1 + \sum_i b_i(s_i' - 1 + \alpha + 1) \\ &< \alpha + 1 + \sum_i b_i(s_i - 1 + \alpha + 1), \end{aligned}$$

which is equivalent to

$$\alpha + \sum_i b_i(s_i' + \alpha) < \alpha + \sum_i b_i(s_i + \alpha),$$

contradicting the optimality of $\mathbf{s}_C(\alpha)$. Thus, we have shown

$$\mathbf{s}_{C'}(\alpha + 1) = (s_1 - 1, \ldots, s_m - 1).$$

Now, follow Algorithm A2 to continue allocating the remaining budget m, and let

$$\mathbf{s}_C(\alpha + 1) := (\tilde{s}_1 - 1, \ldots, \tilde{s}_m - 1)$$

be the result. Clearly, we have $\tilde{s}_i \geqslant s_i$ for all i because the greedy allocation only adds to the units that are already allocated. Since

$$\begin{aligned} u_C(\alpha + 1) &= \alpha + 1 + \sum_i b_i(\tilde{s}_i - 1 + \alpha + 1) \\ &= \alpha + 1 + \sum_i b_i(\tilde{s}_i + \alpha) \end{aligned}$$

we have

$$\begin{aligned} \Delta u_C(\alpha) &:= u_C(\alpha + 1) - u_C(\alpha) \\ &= \sum_i [b_i(\tilde{s}_i + \alpha) - b_i(s_i + \alpha)] + 1 \\ &= \sum_i [\Delta b_i(\tilde{s}_i - 1 + \alpha) + \cdots + \Delta b_i(s_i + \alpha)] + 1. \end{aligned}$$

Recall that $\Delta b_i(\cdot)$ is nondecreasing. So, for any fixed C, $\Delta u_C(\alpha)$ is nondecreasing in α, implying that $u_C(\alpha)$ is convex.

ACKNOWLEDGMENTS

The first author was supported in part by NSF grants DMI-9896339 and DMI-0084922.

The second author was supported in part by NSF grant DMI-0085124. Part of this author's research was undertaken while he was on leave at the Chinese University of Hong Kong, Dept. of Systems Engineering and Engineering Management, and supported by a Direct Grant from CUHK and a HK/RGC grant CUHK4376/99E.

REFERENCES

Agrawal, N., M. Cohen. 2001. Optimal material control and performance evaluation in an assembly environment with component commonality. *Naval Res. Logist.* **48** 409–429.

Buzacott, J. A., J. G. Shanthikumar. 1994. *Stochastic Models of Manufacturing Systems.* Prentice-Hall, Englewood Cliffs, NJ.

Chang, C. S., D. D. Yao. 1990. Rearrangement, majorization, and stochastic scheduling, *Math. Oper. Res.* **18** 658–684.

Cheung, K. L., W. Hausman. 1995. Multiple failures in a multi-item spare inventory model. *IIE Trans.* **27** 171–180.

Clark, A. J., H. Scarf. 1960. Optimal policies for a multi-echelon inventory problem. *Management Sci.* **6** 475–490.

Connors, D. P., D. D. Yao. 1996. Methods for job configuration in semiconductor manufacturing. *IEEE Trans. Semiconductor Manufacturing* **9** 401–411.

Ettl, M., G. E. Feigin, G. Y. Lin, D. D. Yao. 2000. A supply network model with base-stock control and service requirements. *Oper. Res.* **48** 216–232.

Falin, G. 1994. The $M^k/G/\infty$ batch arrival queue with heterogeneous dependent servers. *J. Appl. Probab.* **31** 841–846.

Federgruen, A., P. Zipkin. 1986. An inventory model with limited production capacity and uncertain demands, I: The average cost criterion, II: The discounted cost criterion. *Math. Oper. Res.* **11** 193–207.

Gallien, J., L. Wein. 2001. A simple and effective component procurement policy for stochastic assembly systems. *Queueing Sys.* **38** 221–248.

Glasserman, P. 1997. Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* **45** 244–257.

——, Y. Wang. 1998. Leadtime-inventory tradeoffs in assemble-to-order systems. *Oper. Res.* **46** 858–871.

Hausman, W. H., H. L. Lee, A. X. Zhang. 1998. Order response time reliability in a multi-item inventory system. *Eur. J. Oper. Res.* **109** 646–659.

Lai, T., H. Robbins. 1976. Maximally dependent random variables. *Proc. National Acad. Sci.* **73** 286–288.

Lee, Y., P. Zipkin. 1992. Tandem queues with planned inventories. *Oper. Res.* **40** 936–947.

Rosling, K. 1989. Optimal inventory policies for assembly systems under random demands. *Oper. Res.* **37** 565–579.

Ross, S. M. 1996. *Stochastic Processes*, 2nd ed. Wiley, New York.

Shaked, M., J. G. Shanthikumar. 1994. *Stochastic Orders and Their Applications*. Academic Press, New York.

Shanthikumar, J. G., D. D. Yao. 1991a. Strong stochastic convexity and its applications, *J. Appl. Probab.* **28** 131–145.

——, ——. 1991b. Bivariate characterization of some stochastic order relations. *Adv. Appl. Probab.* **23** 642–659.

Sherbrooke, C. C. 1992. *Optimal Inventory Modeling of Systems*. Wiley, New York.

Song, J. S. 1998. On the order fill rate in a multi-item, base-stock inventory system. *Oper. Res.* **46** 831–845.

——. 2000. A note on assemble-to-order systems with batch ordering. *Management Sci.* **46** 739–743.

——. 2002. Order-based backorders and their implications in multi-item inventory systems. *Management Sci.* **48** 499–516.

——, S. Xu, B. Liu. 1999. Order fulfillment performance measures in an assembly-to-order system with stochastic leadtimes. *Oper. Res.* **47** 131–149.

Wang, Y. 1999. Near-optimal base-stock policies in assemble-to-order systems under service levels requirements. Working paper, MIT Sloan School, Cambridge, MA.

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

Xu, S. H., H. Li. 2000. Majorization of weighted trees: A new tool to study correlated stochastic systems. *Math. Oper. Res.* **25** 298–323.

Zhang, A. X. 1997. Demand fulfillment rates in an assemble-to-order system with multiple products and dependent demands. *Production Oper. Management* **6** 309–324.

Zipkin, P. 2000. *Foundations of Inventory Management*. McGraw-Hill, New York.