LEADTIME-INVENTORY TRADE-OFFS IN ASSEMBLE-TO-ORDER SYSTEMS

PAUL GLASSERMAN

Columbia Business School, New York, New York

YASHAN WANG

MIT Sloan School, Cambridge, Massachusetts

(Received May 1996; revision received December 1996; accepted February 1997)

This paper studies the trade-off between inventory levels and the delivery leadtime offered to customers in achieving a target level of service. It addresses the question of how much a delivery leadtime can be reduced, per unit increase in inventory, at a fixed fill rate. We show that for a class of assemble-to-order models with stochastic demands and production intervals there is a simple *linear* trade-off between inventory and delivery leadtime, in a limiting sense, at high fill rates. The limiting slope is easy to calculate and can be interpreted as the approximate marginal rate for trading off inventory against leadtime at a constant level of service. We also investigate how various model features affect the trade-off—in particular, the impact of orders for multiple units of a single item and of orders for multiple units of different items.

In the production and distribution of goods, inventory is the currency of service. An increase in service can virtually always be achieved through an increase in safety stocks, so a supplier inevitably faces a trade-off between service levels and inventory costs. This and related trade-offs are discussed at least qualitatively in most operations management textbooks (e.g., Section 14-5 of Chase and Aquilano 1995, Chapter 7 of Bowersox and Cooper 1992, Sections 14-4 and 14-5 of Mclain et al. 1992), in some of the managerial literature (e.g., Chapter 5 of Heskett et al. 1990, Zipkin 1991), and in the research literature (e.g., Buzacott and Shanthikumar 1994, Chang 1985, Ettl et al. 1995, Muckstadt and Thomas 1980 and 1983, Schraner 1996, Song 1997, Song et al. 1997). It raises the question of just how much service inventory can buy; i.e., what is the marginal cost of a service improvement, in units of inventory, if the improvement is achieved through increases in inventory?

This is the main question we examine. There is no universal answer, but we show that for a particular class of models there is a simple *linear* trade-off between service and inventory, in a limiting sense, at high levels of service. We identify the limiting slope and interpret it as the approximate marginal rate for trading off inventory against service. We also investigate how other model features affect this relation—in particular, the impact of orders for multiple units of a single item and of orders for multiple different items. Our general approach to identifying the trade-off is applicable in other settings as well.

The models we consider have the following general features. Items are made to stock to supply variable demands for finished products; multiple finished products are assembled-to-order from the items. Service is measured by the *fill rate*, defined here to be the proportion of orders filled within a target interval, called the *delivery leadtime*, or simply the *leadtime*. The system operates under a continuous-review base-stock policy under which each demand for a unit of an item triggers a replenishment order for that item. Items are produced one at a time on dedicated facilities; production intervals may be constant or variable. Reducing the delivery leadtime while maintaining a fill rate of, say, 98% requires increasing inventories, and this is the trade-off we investigate.

Though we focus on fill rates and leadtimes, it is worth noting that in a closely related discrete-time version of the single-item case of our model, a base-stock policy has been shown to minimize holding and backorder costs in Federgruen and Zipkin (1986ab). Glasserman (1997), Liu (1995), and Tayur (1993) discuss the optimal base-stock level. Section 6 of the survey of van Houtum et al. (1995) discusses capacitated models generally. The models of Lee and Zipkin (1992), Veatch and Wein (1994), Zipkin (1986), and the assembly systems of Chapter 6 of Zhang (1996) are also relevant in that they combine features of queueing and inventory systems and (in some cases) study control rules similar to base-stock policies. Extending the results of Clark and Scarf (1962), Rosling (1989) identifies the optimal policy in uncapacitated multistage assembly systems; this is a base-stock policy in the absence of fixed order costs. Ettl et al. (1995) use base-stock policies as approximations to study trade-offs between service levels and inventory costs in a large-scale model of an IBM supply chain. More recently, Schraner (1996) approximates the fill rate in a discrete-time model related to ours and examines trade-offs between *capacity* and inventory, building in part on Hausman et al. (1993). Song (1997) calculates the exact off-the-shelf order fill rate in a

Subject classifications: Production-inventory, assemble-to-order, fill rate, trade-off, exponential tail probability, delivery leadtime.
Area of review: MANUFACTURING OPERATIONS.

0030-364X/98/4606-0858 $05.00
© 1998 INFORMS

related setting, under the assumptions of Poisson demand and deterministic production times. Song et al. (1997) extend the model of Song (1997) to allow for random (exponential) production times. They also study the effect of dependent demands for different items, but the batch sizes are restricted to 0 or 1. Because we allow multiple units of multiple items to be combined into multiple products, our results are also relevant to the literature on component commonality; see, e.g., Baker et al. (1986).

To give some indication of the solutions we arrive at, we begin by describing the simplest possible setting. Orders for individual units of a single item arrive in a Poisson stream with rate λ. The time to produce an individual unit is exponentially distributed with mean $1/\mu$, with $\lambda < \mu$. Let s denote the base-stock level and let x denote the delivery leadtime. Let R denote the steady-state response time of an order—the time elapsed from when an order arrives until it is filled. This is 0 if the order is met from on-hand inventory, and strictly positive otherwise.

Through an evident connection with the M/M/1 queue, the distribution of R is easily found in closed form. The fill rate (the proportion of orders whose response time does not exceed x) is 1 minus

$$P(R > x) = \left(\frac{\lambda}{\mu}\right)^s \exp\{-(\mu - \lambda)x\}. \tag{1}$$

Consider the set of points in (x, s) space for which the expression on the right equals δ, thus resulting in a fill rate of $1 - \delta$, with $0 < \delta < 1$. So long as x and s are both positive, these points are characterized by the relation

$$s = \frac{\log \delta}{\log(\lambda/\mu)} - \frac{\mu - \lambda}{\log(\mu/\lambda)}\, x.$$

In other words, *the level curves of constant service are straight lines*, with slope $-b = -(\mu - \lambda)/\log(\mu/\lambda)$. Decreasing the leadtime by Δx without reducing the fill rate entails increasing s by $b\Delta x$; increasing s to $s + 1$ buys a reduction of $1/b$ in x.

Once we move beyond this Poisson-exponential setting, the simplicity of (1) is lost and the level curves are no longer exactly straight lines. Nevertheless, we show that the trade-off between x and s is approximately linear at high fill rates. The Poisson-exponential setting does not provide much insight into the appropriate slope b in the general case, in part because it contains just two parameters λ and μ. Indeed, this setting could even be considered misleading because substituting the mean interarrival time and mean service time for $1/\lambda$ and $1/\mu$ in the general case is far from correct. We will see, however, that the appropriate generalization of b is quite easily characterized.

The Poisson-exponential setting is useful in providing a first look at the effect of compound demands. Suppose each Poisson arrival brings an order for geometrically many units, with a mean of $1/(1 - p)$, for some $0 < p < 1$ with $\lambda < \mu(1 - p)$. An order is filled only when all units demanded have been provided. The time to produce a geometric batch is the sum of geometrically many exponential random variables and is thus again exponential; however, this setting cannot be entirely reduced to the previous one because the inventory level at an arrival may be strictly positive but insufficient to fill the order. The system is still tractable and results in

$$P(R > x) = \left(\frac{p\mu + \lambda}{\mu}\right)^s \exp\{-((1 - p)\mu - \lambda)x\}.$$

Thus, the level curves of constant service are still straight lines in the positive (x, s) orthant. The new slope $[(1 - p)\mu - \lambda]/\log(\mu/(p\mu + \lambda))$ reflects the effect of batch arrivals. This, too, generalizes, but not in a way that is obvious from the formula alone.

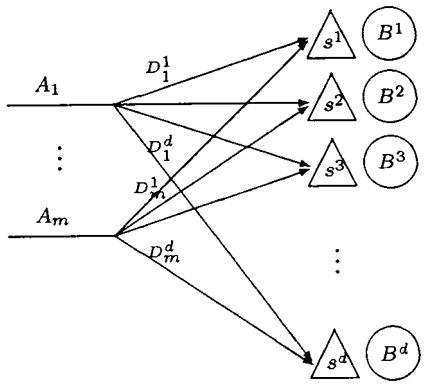

Figure 1. Multiple items assembled to order into multiple products. The demand process for product j is independent of those for other products and has generic interarrival time A_j. An order for product j requires a generic item portfolio: D_j^l units of item $1, \ldots,$ D_j^d units of item d. Item i has generic production time B^i.

The rest of this paper is organized as follows. The next section describes the models we consider in more detail—particularly the extension to multiple items and multiple combinations of items—and then formulates our main results. Section 2 illustrates our results numerically and discusses further approximations. Sections 3–4 develop the necessary tools to prove our main results. Section 5 contains some concluding remarks.

1. MAIN RESULTS

1.1. Model Details and Notation

Figure 1 illustrates the general setting we consider. Multiple items are produced on dedicated facilities (the circles) and kept in inventories (the triangles). A *product* is a collection of a possibly random number of items of each type. It may be convenient to think of the items as components that are assembled into products. When we treat assembly times explicitly (see the discussion after Theorem 1) we model them as random delays—i.e., the assembly operation is uncapacitated. Since this additional delay has little impact on our results, we omit it from most of our discussion.

We use the following notation, often modified by subscripts and superscripts:

A = order interarrival time;
B = unit production interval;
D = batch order size;
R = response time;
s = base-stock level;
x = delivery leadtime.

A superscript i refers to item i, ranging from 1 to d. Most of the paper considers single product systems, where a subscript n refers to the nth order. For example, D_n^i is the number of units of item i required by the nth order; R_n^i is the time taken to fill the nth order's demand for item i, and R_n is the time taken to fill the nth order completely. Occasionally, we need to use double subscripts, as in $B_{n,j}^i$, which denotes the production time for the jth unit of item i in the nth order. An R without a subscript refers to a steady-state response time. Production intervals, interarrival times, and batch sizes are all independent of each other, but we allow dependence among the batches of different items required by an order. In other words, the vectors $\{(D_n^1, \ldots, D_n^d), n = 1, 2, \ldots\}$ are i.i.d., but their components may be dependent. (This is particularly important if different groups of items constitute different products.) When there are multiple products, a subscript j refers to product j, ranging from 1 to m. To simplify the formulation of our results, we assume throughout that the interarrival times have a continuous distribution.

For any random variable Y, the symbol ψ_Y denotes the function

$$\psi_Y(\theta) = \log \mathsf{E}[e^{\theta Y}], \tag{2}$$

called the *cumulant generating function* (c.g.f.) of Y. The function ψ_Y is convex, and it is differentiable in the interior of its domain (the set of θ at which it is finite). The c.g.f.s of all our input random variables will be finite for some $\theta > 0$, and this implies $\psi_Y'(0) = \mathsf{E}[Y]$ and $\psi_Y''(0) = \mathsf{Var}[Y]$. See Chapter 3 of Kendall (1987) for relevant background.

1.2. The Trade-Offs

We begin with a result for the case of a single item. In this setting, we may omit the superscript i. We require that $\mathsf{E}[D]\mathsf{E}[B] < \mathsf{E}[A]$ so that a steady-state response time exists. Defining $X = \sum_{j=1}^{D} B_j - A$, we have

$$\psi_X(\theta) = \psi_D(\psi_B(\theta)) + \psi_A(-\theta). \tag{3}$$

In the following result, the notation $\lim_{s+x\to\infty}$ refers to the limit as either $s \to \infty$ (through integer values), or $x \to \infty$, or both.

Theorem 1. *If there is a $\gamma > 0$ at which $\psi_X(\gamma) = 0$, then with $\beta = \psi_B(\gamma)$,*

$$\lim_{s+x\to\infty} e^{\gamma x + \beta s} P(R(s) > x) = C \tag{4}$$

for some constant $C > 0$.

Based on this result, we interpret $-\gamma/\beta$ as the approximate slope of the trade-off between s and x at high fill rates. To see why, notice that (4) suggests the approximation $P(R(s) > x) \approx C \exp(-\gamma x - \beta s)$. A level curve of constant service is a set of (x, s) points for which $P(R(s) > x) = \delta$, for some $0 < \delta < 1$, and this is given approximately by the set of solutions to $C \exp(-\gamma x - \beta s) = \delta$. In the positive (x, s) orthant, these solutions form the line

$$s = -\frac{\gamma}{\beta} x + \frac{1}{\beta} \log(C/\delta).$$

Thus, a leadtime reduction of Δx entails an increase in s of $(\gamma/\beta)\Delta x$, and a one-unit increase in s buys a reduction of β/γ in x, if the fill rate is to remain unchanged. It is not hard to verify that the ratio γ/β specializes to the ratios given in the previous section in the Poisson-exponential example.

We can extend the setting in Theorem 1 to include an explicit assembly time U_n for the nth order after all the components it requires are available, with U_n i.i.d. and bounded. We assume no congestion and no finished goods inventory at the assembly stage, so U_n acts as an extra delay in the response time. This changes only the constant C in (4), as we explain after the proof of Theorem 1 in Section 3.

If γ in Theorem 1 exists, it is unique because ψ_X is convex and $\psi_X(0) = 0$. Sufficient conditions for the existence of γ are that $\psi_X(\theta)$ be finite for some $\theta > 0$ and that $\psi_X(\theta)$ not jump to infinity as θ increases—i.e., if $\bar{\theta} = \sup\{\theta \geq 0 : \psi_X(\theta) < \infty\}$, then $\psi_X(\theta) \to \infty$ as $\theta \uparrow \bar{\theta}$. These conditions are met for virtually all commonly used distributions. (An exception is the lognormal distribution, which has no exponential moments.) To avoid repeating technical conditions, we assume throughout that all input random variables (A, B, and D) satisfy these conditions.

In general, only a partial characterization of the constant C in (4) is available. However, in the important special case of Poisson arrivals, we obtain an explicit formula:

Proposition 1. *In the setting of Theorem 1, if arrivals are Poisson with rate λ, then*

$$C = \lambda^{-1}(\lambda + \gamma)(1 - \lambda\mathsf{E}[D]\mathsf{E}[B]) \cdot (\psi_D'(\beta)\psi_B'(\gamma)(\lambda + \gamma) - 1)^{-1}.$$

Consider, next, a system with d items but a single product and thus just one arrival stream. Each item i has a base-stock level s^i, so we need to make an assumption about how these scale to get a limiting result. Let $s = s^1 + \cdots + s^d$ and $k_i = s^i/s$, $i = 1, \ldots, d$; we hold these ratios constant as s increases. This assumes that the proportion of total inventory held in each item remains constant, though we could just as easily assume that, e.g., the proportion of work content or holding cost for each item remains constant; this would merely change the constants k_i in the subsequent analysis. For each item i define X^i from A, B^i, and D^i paralleling the definition of X just before (3) and set

$$\psi_i(\theta) = \psi_{D^i}(\psi_{B^i}(\theta)) + \psi_A(-\theta), \tag{5}$$

in analogy with (3). Clearly, Theorem 1 applies to each item separately. Suppose $\gamma_i > 0$ solves $\psi_i(\gamma_i) = 0$ and set $\beta_i = \psi_{B^i}(\gamma_i)$, $\alpha_i = k_i\beta_i$. Then Theorem 1 implies

$$\lim_{s+x\to\infty} e^{\gamma_i x+\alpha_i s}P(R^i(s) > x) = C_i$$

for some $C_i > 0$. The response time for the full order is the maximum of the response times for the individual items required. Its behavior is a bit more subtle because of the interactions among the multiple items.

Let $\gamma = \min_i \gamma_i$ and $\mathcal{I}_x = \{i:\gamma_i = \gamma\}$; these are the set of *leadtime-critical* items in the sense that their individual fill rates increase most slowly as x increases to ∞. Let $\alpha = \min_i \alpha_i$ and $\mathcal{I}_s = \{i:\alpha_i = \alpha\}$; these are similarly the set of *inventory-critical* items because their fill rates increase most slowly as s increase to ∞. These sets of items determine the product fill rate when x or s becomes large. To exclude trivial cases, we assume that for any two items i and j in $\mathcal{I}_x$ or $\mathcal{I}_s$, we have $P(X^i \neq X^j) > 0$. Indeed, the only case in which this fails is if the two items are always ordered in the same quantity and take the same time to produce, in which case they should be modeled as a single item.

Theorem 2. *Suppose the solutions* $\gamma_1, \ldots, \gamma_d$ *all exist. Then*

$$\lim_{x\to\infty} e^{\gamma x}P(R(s) > x) = \sum_{i\in\mathcal{I}_x} C_i e^{-\alpha_i s}, \tag{6}$$

and if either $|\mathcal{I}_s| = 1$ *or* $\{D^i, i \in \mathcal{I}_s\}$ *is independent, then*

$$\lim_{s\to\infty} e^{\alpha s}P(R(s) > x) = \sum_{i\in\mathcal{I}_s} C_i e^{-\gamma_i x}. \tag{7}$$

The condition that the D^i be independent is far from necessary for (7), even if $|\mathcal{I}_s| > 1$. From the proof of Theorem 2 it will be clear that (7) holds under the much weaker condition (35), and it may well be true without even this additional condition.

If $\mathcal{I}_x = \mathcal{I}_s = \mathcal{I}$, Theorem 2 yields (without assuming independence or (35))

$$\lim_{s+x\to\infty} e^{\gamma x+\alpha s}P(R(s) > x) = \sum_{i\in\mathcal{I}} C_i \equiv C_{\mathcal{I}},$$

which provides a simpler counterpart to Theorem 1 and suggests the approximation

$$P(R(s) > x) \approx C_{\mathcal{I}} e^{-\gamma x-\alpha s}.$$

It is only in this case that we can give $-\gamma/\alpha$ the simplest interpretation of the slope of the trade-off between x and s. In the general case, $\mathcal{I}_x$ and $\mathcal{I}_s$ represent the sets of items that constrain the fill rate at long delivery intervals and high base-stock levels, respectively. When these are not the same, different trade-offs apply in different regions. This is further explored in Section 2.

The final variant we consider allows multiple sets of items to be combined into m distinct products. In this setting, we require that arrivals of orders for the various products follow independent (compound) Poisson processes. We also need to vary our notation slightly to distinguish products from items: we use subscripts for products and continue to use superscripts for items. Orders for product j arrive at rate λ_j, and each order of product j requires D_j^i units of item i.

Let $\mathcal{I}_j = \{i:P(D_j^i > 0) > 0\}$ be the set of items required by product j; let $\mathcal{P}^i = \{j:P(D_j^i > 0) > 0\}$ be the set of products requiring item i. For each item i, the demand is the superposition of independent (compound) Poisson processes with $\lambda^i = \sum_{j\in\mathcal{P}^i} \lambda_j$; the batch size D^i is distributed as a mixture of $\{D_j^i\}$; i.e., with probability λ_j/λ^i, D^i is distributed as D_j^i, for $j \in \mathcal{P}^i$. With γ^i and α^i calculated just as before, Theorem 1 applies to R^i, the steady-state item-i response time. Let R_j be the steady-state response time for product j; then $R_j = \max_{i\in\mathcal{I}_j} R^i$. Define

$$\bar{\gamma}_j = \min_{i\in\mathcal{I}_j}\{\gamma^i\} \quad \text{and} \quad \mathcal{I}_j^x = \{i \in \mathcal{I}_j : \gamma^i = \bar{\gamma}_j\};$$

$\mathcal{I}_j^x$ is the set of leadtime-critical items for product j. Also define

$$\bar{\alpha}_j = \min_{i\in\mathcal{I}_j}\{\alpha^i\} \quad \text{and} \quad \mathcal{I}_j^s = \{i \in \mathcal{I}_j : \alpha^i = \bar{\alpha}_j\};$$

$\mathcal{I}_j^s$ is the set of inventory-critical items for product j. We now have Theorem 3.

Theorem 3. *Suppose the solutions* $\gamma^1, \ldots, \gamma^d$ *all exist. Then*

$$\lim_{x\to\infty} e^{\bar{\gamma}_j x}P(R_j(s) > x) = \sum_{i\in\mathcal{I}_j^x} C_i e^{-\alpha^i s}, \tag{8}$$

and if $|\mathcal{I}_j^s| = 1$ *or* $\{D_j^i, i \in \mathcal{I}_j^s\}$ *are independent, then*

$$\lim_{s\to\infty} e^{\bar{\alpha}_j s}P(R_j(s) > x) = \sum_{i\in\mathcal{I}_j^s} C_i e^{-\gamma^i x}. \tag{9}$$

The same comments on the independence condition for the Ds after Theorem 2 also apply here. As with Theorem 2 the cleanest version of this result applies when $\mathcal{I}_j^x = \mathcal{I}_j^s$; i.e., the leadtime-critical and inventory-critical items coincide. Because the result has been specialized to the case of Poisson arrivals, Proposition 1 applied to each item i yields an expression for each C_i.

2. APPLYING THE APPROXIMATION

In this section, we interpret and test the results of Section 1. We first illustrate the trade-offs in single-item systems through several examples. Then, we discuss further approximations for the trade-off parameters γ and β. Finally, we address issues in applying the trade-off when there is interaction among multiple items.

2.1. Single-Item Systems

Theorem 1 suggests that when the fill rate is high, varying the values of leadtime x and inventory level s according to the linear trade-off rule

$$\Delta s = -\frac{\gamma}{\beta}\Delta x \tag{10}$$

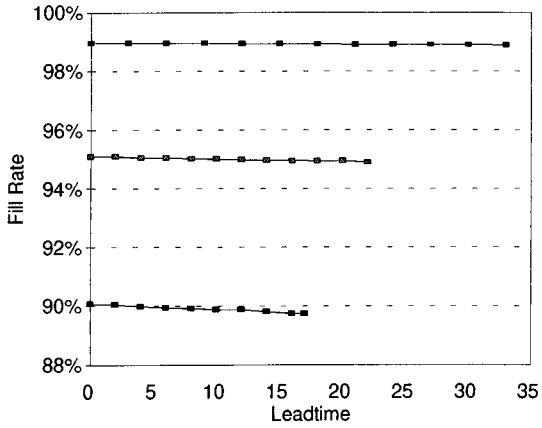

Figure 2. Compound Poisson demand process, $\mathsf{E}[A] = 2.4$; $P_D(1) = 0.7$, and $P_D(4) = 0.3$; Erlang production time with $c_B^2 = 0.2$; $\rho = 80\%$.

entails little change in the fill rate (see the discussion following Theorem 1). Indeed, the smaller the resulting change in the fill rate, the better the linear approximation to the trade-off. We will test how the linear approximation works through several examples. In each example, we first calculate γ and β according to Theorem 1. We study the systems from $x = 0$ and choose some $s > 0$ such that the actual fill rate is high (90% or higher). We then make a series of increases in x and decrease s according to the trade-off rule (10) until s drops to 0. This way we get a series of (x, s) pairs. At each pair the actual fill rate is estimated by Monte Carlo simulation. Plotting the fill rate against the leadtime x yields a curve; when this curve is nearly a horizontal line the linear approximation in (10) works well.

The procedure to estimate the fill rate is discussed after the proof of Theorem 1 in the next section. We will test several commonly used distributions at different utilization levels, where by utilization we mean the ratio of the mean production time of a random batch to the mean interarrival time. By choosing appropriate s values at $x = 0$, we

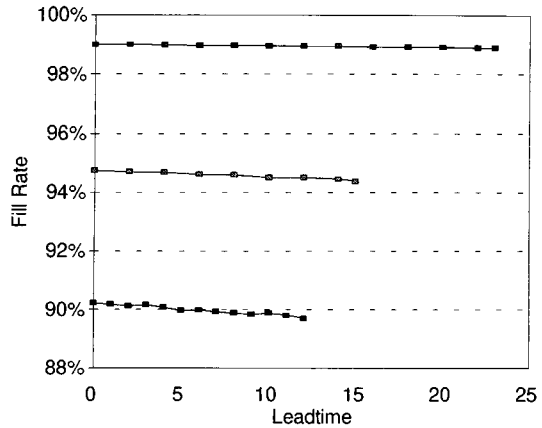

Figure 3. Compound Poisson demand process, $\mathsf{E}[A] = 2.7$; $P_D(1) = 0.7$, and $P_D(4) = 0.3$; Erlang production time with $c_B^2 = 0.2$; $\rho = 70\%$.

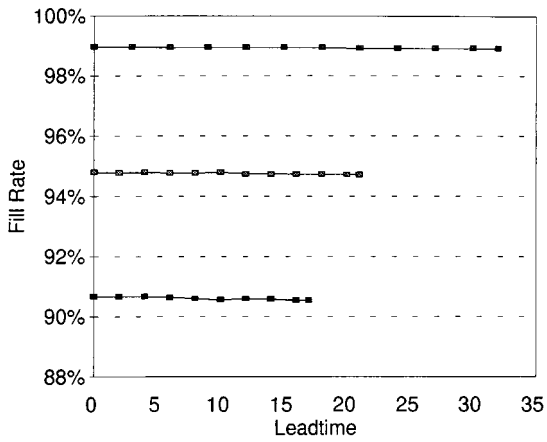

Figure 4. Compound Poisson demand process, $\mathsf{E}[A] = 2.4$; $P_D(1) = 0.7$, and $P_D(4) = 0.3$; normally distributed production time with $\mathsf{Var}[B] = 0.9$; $\rho = 80\%$.

study the trade-offs when the fill rate is around 90%, 95%, and 99%. (At higher service levels, s has a larger value at $x = 0$ and reaches 0 at a larger x, so the resulting curve is longer.) The simulation results are graphed in Figures 2–7, with three curves corresponding to the three service levels in each figure. The captions specify the distributions of A, B, and D; all cases have $\mathsf{E}[B] = 1$. We use c_A and c_B to denote the coefficient of variation for A and B, respectively, and use $P_D(n)$ to denote $P(D = n)$ and ρ to denote the system utilization.

From the examples, we make the following observations.

1. In all the examples studied the simulated fill rate curves are very close to flat, straight lines, regardless of the difference in distributions and utilization levels. This means that varying x and s according to the linear s-x trade-off rule (10) indeed yields approximately the same fill rate. Hence the linear limiting trade-off between x and s captures the essence of the relation between x and s.

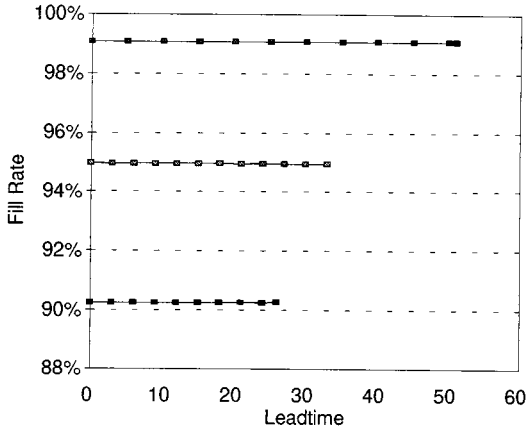

Figure 5. Compound Poisson demand process, $\mathsf{E}[A] = 2.5$; $P_D(1) = P_D(2) = P_D(3) = P_D(4) = 0.1$; deterministic production time; $\rho = 90\%$.

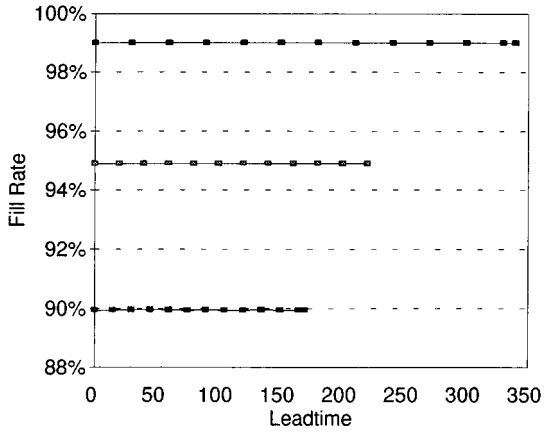

Figure 6. Compound Poisson demand process, $\mathsf{E}[A] = 2.24$; $P_D(1) = P_D(2) = P_D(3) = 0.3$; deterministic production time; $\rho = 98\%$.

2. The approximation works very well when the service level is reasonably high, and the higher the fill rate, the better the results we get. When the fill rate is around 99%, the linear trade-off rule is virtually exact. This is consistent with the fact that at higher service level the system is closer to the limit regime of Theorem 1.

3. Along each fill rate curve in the figures above, x and s vary over a very wide range (in fact, over the entire possible range from $x = 0$ to $s = 0$), yet the curves remain nearly flat and straight. Thus, the trade-off rule (10) is not just a local property. Changes in delivery leadtime and inventory level, Δx and Δs, do not have to be very small; they can be made to be very big as long as $x + \Delta x$ and $s + \Delta s$ remain nonnegative.

4. We observe a decline of the 90% fill rate curve as x increases in Figures 2 and 3. This is mainly due to a rounding effect: for some changes in x, the corresponding change in s given by (10) may not be an integer. But s has to be integral because of its physical meaning, so we round the nonintegral change to the nearest integer. In these examples, it happens that we always round down so the fill rate curves go down. We do not see the decline of the fill rate in Figures 4 and 5, where rounding is not necessary.

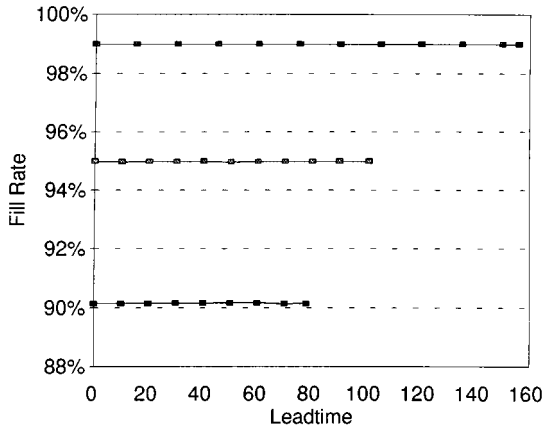

Figure 7. Hyperexponential interarrival time, $\mathsf{E}[A] = 2.5$; $c_A^2 = 4$; $P_D(1) = P_D(2) = P_D(3) = 0.3$; $P_D(4) = 0.1$; deterministic production time; $\rho = 88\%$.

2.2. Two-Moment Approximation

Calculating the trade-off parameters γ and β requires knowledge of the distributions of A, B and D, which may not always be available. If we only have partial knowledge of the distributions—specifically, the means and variances—we approximate $\gamma > 0$ through a two-moment approximation for the ψ_X in (3), i.e., we set $\psi_X(\theta) \approx \mathsf{E}[X]\theta + (1/2)\mathsf{Var}[X]\theta^2 = 0$ and solve to get

$$\gamma \approx -\frac{2\mathsf{E}[X]}{\mathsf{Var}[X]}, \tag{11}$$

where

$$\mathsf{E}[X] = \mathsf{E}[B]\mathsf{E}[D] - \mathsf{E}[A]$$

and

$$\mathsf{Var}[X] = \mathsf{E}[D]\mathsf{Var}[B] + \mathsf{Var}[D](\mathsf{E}[B])^2.$$

Similarly, by the two-moment approximation for ψ_B we get

$$\beta = \psi_B(\gamma) \approx E[B]\gamma + \frac{1}{2}\,\mathsf{Var}[B]\gamma^2. \tag{12}$$

We tested the linear trade-off of (10) on the same systems studied in the previous subsection, but replacing γ and β with their two-moment approximations (11) and (12). The resulting graphs are virtually indistinguishable from the previous ones and are therefore omitted. This indicates that the two-moment approximation is adequate in practice.

2.3. Multiple-Item Systems

In Section 1 we stated two limiting results (6) and (7) for the tail probability of the product response time, one on x and one on s. When the leadtime x is long, $P(R > x) \approx \sum_{i \in \mathcal{J}_x} C_i e^{-\gamma x - \beta_i s^i}$ and the product fill rate is constrained by the items with the smallest γ (leadtime-critical items). When the total inventory level s is high, $P(R > x) \approx \sum_{i \in \mathcal{J}_s} C_i e^{-\gamma_i x - \alpha s}$ and the product fill rate is constrained by the items with the smallest α (inventory-critical items). While it is sometimes possible to determine which regime applies, in other cases, the dominating effect of the items in $\mathcal{J}_x$ or $\mathcal{J}_s$ may not be evident, since x and s are always finite in practice. We address this issue next.

When product fill rate is high, the fill rate of each item i must be as high or higher, and is approximately equal to $1 - C_i e^{-\gamma_i x - \beta_i s^i}$. Obviously, items with relatively small fill rates are the ones that constrain the product fill rate. So we propose the following criterion to determine a set $\mathcal{J}$ of constraining items. For each item i we calculate $\hat{p}_i \triangleq e^{-\gamma_i x - \beta_i s^i}$ as a surrogate for $P(R^i(s) > x)$ and take $\mathcal{J}$ to be the set of items with high $\hat{p}_i$. (The constant C_i is, of course, unknown in general.) When the leadtime x is changed, inventory levels of the items in $\mathcal{J}$ should be varied according to the item-level trade-off rule

$$\Delta s^i = -\frac{\gamma_i}{\beta_i}\,\Delta x \tag{13}$$

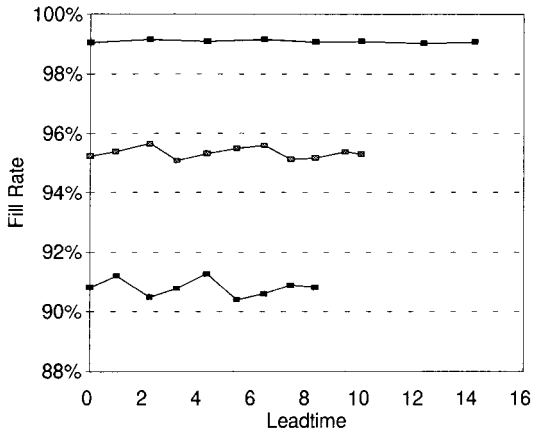

Figure 8. Poisson demand process, $\mathsf{E}[A] = 2.4$; $P_D(1, 1) = 0.4$; $P_D(1, 2) = 0.3$; B^1, B^2 have Erlang distribution, $\mathsf{E}[B^1] = 1$, $c^2_{B_1} = 0.2$, $\mathsf{E}[B^2] = 0.8$, $c^2_{B_1} = 0.2$; $\rho_2 = 63\%$.

to maintain the same product fill rate. Our experience from numerical studies is that when $\Sigma_{i\in\mathcal{J}}\ \hat{p}_i/\Sigma_{i=1}^{d}\ \hat{p}_i \geqslant 80\%$, we get satisfactory results, as will be illustrated through several examples.

We first study two-item systems with $\hat{p}_1$ and $\hat{p}_2$ close to each other so $\mathcal{J}$ contains both items. Much as in the single item case, we choose some $s^1 > 0$, $s^2 > 0$ at $x = 0$ such that the product fill rate is high (90% or higher). We then make a series of increases in x and decrease both s^1 and s^2 according to the item-level trade-off (13) until either s^i drops to 0. By estimating the product fill rate at each (x, s^1, s^2) triple we get a plot of fill rate against leadtime x. Again, a horizontal curve means the linear s-x trade-off relation and the proposed mechanism to identify constraining items work well. The results of the two-item systems are in Figures 8–13. In the captions, we use $P_D(m, n)$ to denote $P(D^1 = m, D^2 = n)$, and ρ_1, ρ_2 to denote the utilization level of items 1 and 2.

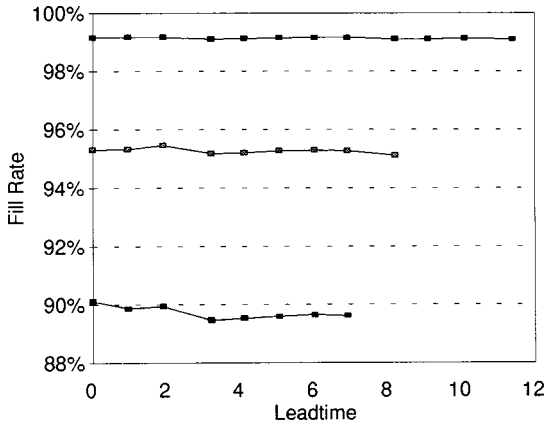

Figure 9. Poisson demand process, $\mathsf{E}[A] = 2.7$; $P_D(1, 1) = 0.4$; $P_D(1, 2) = 0.3$; $P_D(4, 3) = 0.3$; B^1, B^2 have Erlang distribution, $\mathsf{E}[B^1] = 1$, $c^2_{B1} = 0.2$, $\mathsf{E}[B^2] = 0.8$, $c^2_{B1} = 0.2$; $\rho_1 = 70\%$; $\rho_2 = 56\%$.

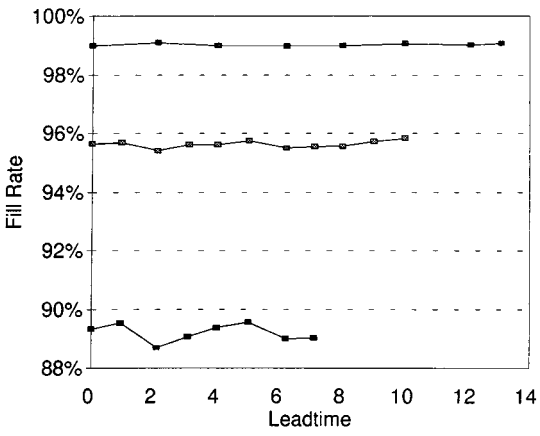

Figure 10. Poisson demand process, $\mathsf{E}[A] = 2.4$; $P_D(1, 1) = 0.4$; $P_D(1, 2) = 0.3$; $P_D(4, 3) = 0.3$; B^1, B^2 have normal distribution, $\mathsf{E}[B^1] = 1$, $\mathsf{Var}[B^1] = 0.09$, $\mathsf{E}[B^2] = 0.8$, $\mathsf{Var}[B^2] = 0.625$; $\rho_1 = 79\%$; $\rho_2 = 63\%$.

Finally, we consider a five-item system with a Poisson order process, having $\mathsf{E}[A] = 3$. The order quantities have the following probabilities: $P_D(2, 3, 4, 2, 1) = 0.2$, $P_D(3, 2, 2, 4, 3) = 0.3$, $P_D(4, 4, 1, 1, 3) = 0.2$ and $P_D(1, 1, 2, 2, 2) = 0.3$. Production times all have Erlang distributions with $c^2 = 0.2$, $\mathsf{E}[B^1] = \mathsf{E}[B^2] = \mathsf{E}[B^3] = 1$, $\mathsf{E}[B^4] = \mathsf{E}[B^5] = 0.9$. We choose different s^i values at $x = 0$ so that different sets of items are binding. The results are illustrated in Figures 14–17. In Figure 14, all five items have similar $\hat{p}$ values so the constraining set $\mathcal{J} = \{1, 2, 3, 4, 5\}$. In Figure 15 one item is constraining and $\mathcal{J} = \{1\}$. In Figures 16 and 17, the same three items are constraining, $\mathcal{J} = \{1, 2, 3\}$. The difference is that in Figure 16 the trade-offs are on the constraining items, whereas in Figure 17 the trade-offs are on all five items—when x changes, s^4 and s^5 also change according to their trade-off equations, although items 4

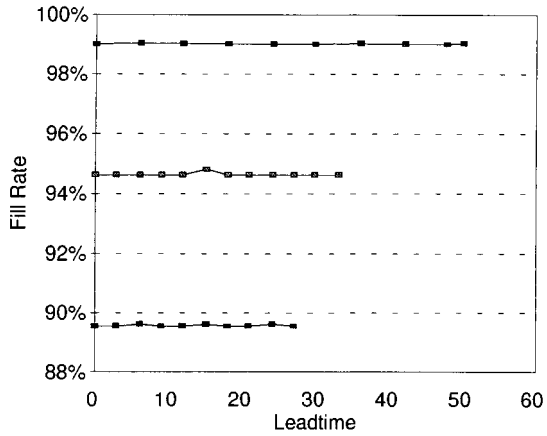

Figure 11. Poisson demand process, $\mathsf{E}[A] = 2.5$; $P_D(1, 2) = P_D(1, 3) = P_D(1, 4) = P_D(2, 2) = P_D(3, 4) = P_D(4, 1) = 0.1$, $P_D(2, 3) = P_D(3, 1) = 0.2$; deterministic production times $B^1 = 1$, $B^2 = 0.9$; $\rho_1 = 88\%$; $\rho_2 = 86\%$.

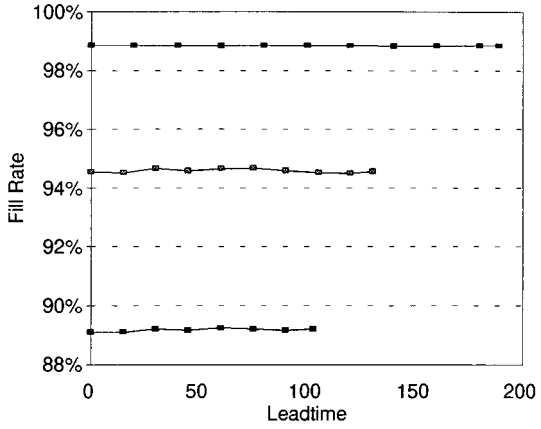

Figure 12. Poisson demand process, $\mathsf{E}[A] = 2.24$; $P_D(1, 2) = P_D(1, 3) = P_D(1, 4) = P_D(2, 2) = P_D(3, 4) = P_D(4, 1) = 0.1$, $P_D(2, 3) = P_D(3, 1) = 0.2$; deterministic production times $B^1 = 1$, $B^2 = 0.9$; $\rho_1 = 98\%$; $\rho_2 = 96\%$.

and 5 are considered nonbinding. This is for comparison with Figure 16.

From the figures, we see that the proposed trade-off rule generally works well, again, regardless of the distribution and utilization level. In some examples, the lowermost (90%) curve is jagged. One of the main reasons is the rounding effect, which is more severe here than in the single-item case. For example, on the 90% curve in Figure 8, the s^2 value should be 4.804 for the fifth point and 3.474 for sixth point according to (13). After rounding, the values become 5 and 3, respectively. A difference of $4.804 - 3.474 = 1.33$ increases to $5 - 3 = 2$, so we see a drop in the fill rate. Two-moment approximations for γ_i and β_i yield virtually the same fill rate graphs.

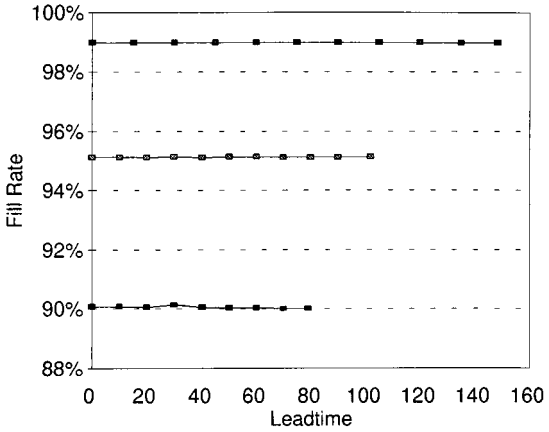

Figure 13. Hyperexponential interarrival time, $\mathsf{E}[A] = 2.5$; $c_A^2 = 4$; $P_D(1, 2) = P_D(1, 3) = P_D(1, 4) = P_D(2, 2) = P_D(3, 4) = P_D(4, 1) = 0.1$, $P_D(2, 3) = P_D(3, 1) = 0.2$; deterministic production times $B^1 = 1$, $B^2 = 0.9$; $\rho_1 = 88\%$; $\rho_2 = 86\%$.

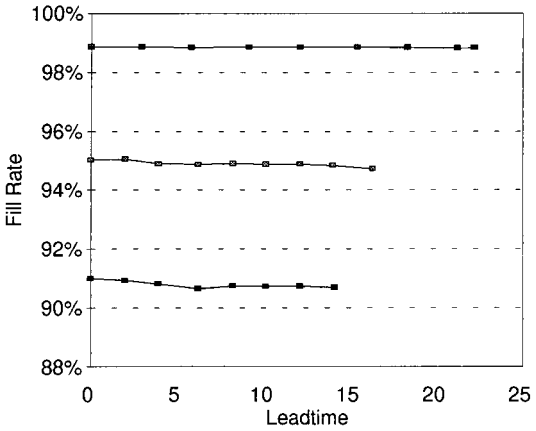

Figure 14. A five-item system. All five items are equally constraining, and trade-offs are on five items.

3. ANALYSIS OF THE SINGLE-ITEM SYSTEM

The main purpose of this section is to prove Theorem 1. A useful first step is a characterization of the response time in terms of an associated queue. Consider, then, a batch-arrival queue with batch interarrival times $\{A_n\}$, batch sizes $\{D_n\}$, and individual production times $\{B_{n,1}, \ldots, B_{n,D_n}\}$. The queue is empty at time zero. The following sample-path result relates the response times $R_n(s)$ in the original system to the waiting times in the associated queue when both systems are driven by the same inputs.

Lemma 1. *Let W_n be the waiting time of batch n in the queue. Then*

$$R_n(s) = \left(W_{N_n} + \sum_{j=1}^{H_n} B_{N_n,j} - \sum_{j=1}^{n-N_n} A_{n-j} \right)^+, \tag{14}$$

where

$$N_n = \sup\left\{ 1 \leq k \leq n : \sum_{j=k}^{n} D_j > s \right\}, \tag{15}$$

and

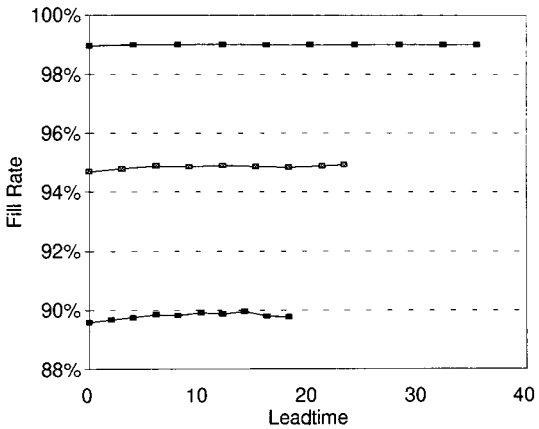

Figure 15. A five-item system. One item is constraining and the trade-off is on that constraining item.

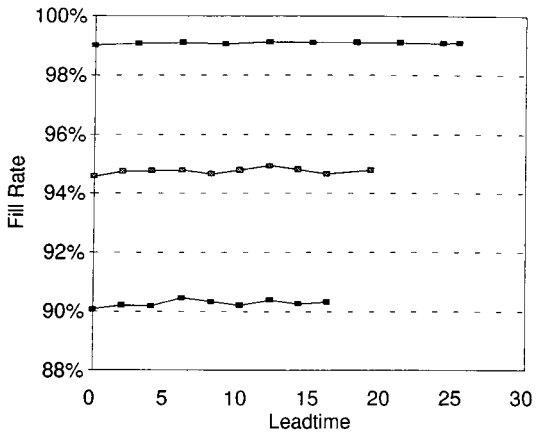

Figure 16. A five-item system. Trade-offs are on the three constraining items.

$$H_n = \sum_{j=N_n}^{n} D_j - s. \tag{16}$$

Proof. In the queue, let $\{t_n, n \geq 1\}$ be the batch arrival epochs, let $\{t'_n, n \geq 1\}$ be the batch departure epochs and let $\{\hat{t}(k), k \geq 1\}$ be the service completion epochs of the individual jobs. In the production-inventory system, the inventory level starts from s at time 0. The cumulative demand at time t_n is $\Sigma_{j=1}^{n} D_j$. For n with $\Sigma_{j=1}^{n} D_j \leq s$, no waiting is necessary for demand n and $R_n(s) = 0$; equation (14) holds trivially because $N_n = 1$, $H_n \leq 0$ and $W_1 = 0$. For n with $\Sigma_{j=1}^{n} D_j > s$,

$$R_n(s) = \left(\hat{t}\left(\sum_{j=1}^{n} D_j - s \right) - t_n \right)^+.$$

To see this, notice that at time $\hat{t}(\Sigma_{j=1}^{n} D_j - s)$ the system finishes producing $\Sigma_{j=1}^{n} D_j - s$ units. With the initial inventory of s units, there are cumulatively $\Sigma_{j=1}^{n} D_j$ units available to meet the first n demands at this time. By the definition of N_n,

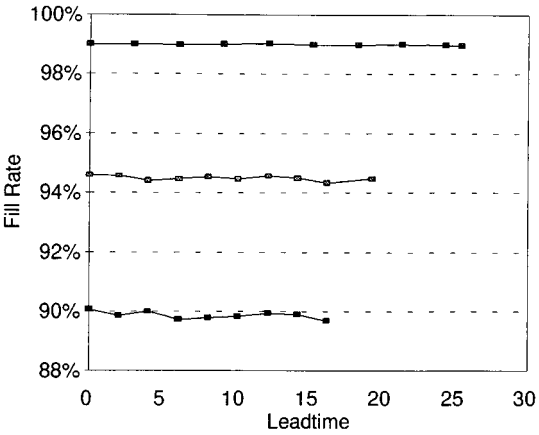

Figure 17. A five-item system. Same three items as in Figure 16 are constraining, but the trade-offs are on all five items for comparison with Figure 16.

$$t'_{N_n} - \sum_{j=1}^{D_{N_n}} B_{N_n,j} < \hat{t}\left(\sum_{j=1}^{n} D_j - s \right) \leq t'_{N_n}.$$

A little careful bookkeeping shows that $\hat{t}(\Sigma_{j=1}^{n} D_j - s) = t'_{N_n} - \Sigma_{j=1}^{D_{N_n}} B_{N_n,j} + \Sigma_{j=1}^{H_n} B_{N_n,j}$ and

$$\hat{t}\left(\sum_{j=1}^{n} D_j - s \right) - t_n = t'_{N_n} - t_{N_n} + t_{N_n} - t_n - \sum_{j=1}^{D_{N_n}} B_{N_n,j} + \sum_{j=1}^{H_n} B_{N_n,j} = W_{N_n} + \sum_{j=1}^{H_n} B_{N_n,j} - \sum_{j=1}^{n-N_n} A_{n-j}.$$

The result (14) follows. □

In words, Equation (14) states that order n is filled when the system finishes producing all the units of order $(N_n - 1)$, plus the first H_n units of order N_n. By construction, W_{N_n} is determined by $\{A_j, (B_{j,1}, \ldots, B_{j,D_j}), D_j; j \leq N_n - 1\}$ and (N_n, H_n) are determined by $\{D_j, N_n \leq j \leq n\}$. Under our independence assumptions, $\{A_j, (B_{j,1}, \ldots, B_{j,D_j}), D_j; j \leq k - 1\}$, and $\{D_j, j \geq k\}$ are independent of each other for any k, and therefore

$$P(W_{N_n} \leq x | N_n = k) = P(W_k \leq x | N_n = k) = P(W_k \leq x), \tag{17}$$

for any x. As $n \to \infty$, $N_n \to \infty$, a.s., and if the system is stable (i.e., $\mathsf{E}[D_1]\mathsf{E}[B_1] - \mathsf{E}[A_1] < 0$) the W_n converge weakly to a random variable W having the distribution of the steady-state batch waiting time in the queue. Using (17) in the first equation of the proof of Theorem 1.1.1 of Gut (1988), we conclude that the W_{N_n} also converge in distribution to W. Indeed, a small extension of Gut's result shows that $(W_{N_n}, n - N_n, H_n)$ converge in distribution to $(W, N - 1, H)$ with

$$N = N(s) \triangleq \min\left\{ k \geq 1 : \sum_{j=1}^{k} \bar{D}_j > s \right\}, \tag{18}$$

and

$$H = H(s) \triangleq \sum_{j=1}^{N(s)} \bar{D}_j - s, \tag{19}$$

and W independent of (N, H). (Here, the $\bar{D}_j$ are i.i.d. with the same distribution as the D_j.) In light of Lemma 1, the steady-state response time can therefore be represented as

$$R(s) = \left(W + \sum_{j=1}^{H(s)} \bar{B}_j - \sum_{j=1}^{N(s)-1} \bar{A}_j \right)^+, \tag{20}$$

where the $\bar{A}_j$ and $\bar{B}_j$ are i.i.d. with the same distributions as the original interarrival and unit production times.

The proof of Theorem 1 uses an exponential change of measure (also called exponential twisting), so we briefly review this concept; see Chapter XII of Asmussen (1987) or Chapter VIII of Siegmund (1985) for additional background. Suppose a random variable Z has distribution F and that

$$\psi_Z(\theta) = \log \int e^{\theta x} \, dF(x)$$

is finite in some nondegenerate interval containing 0. Then

$$dF_{(\theta)}(x) = e^{\theta x - \psi_Z(\theta)}\, dF(x)$$

defines a family of distributions indexed by θ. Let $P_{(\theta)}$ be the probability measure under which Z has distribution $F_{(\theta)}$; we say that $P_{(\theta)}$ is obtained by θ-twisting Z. For any integrable function g, the expectation of $g(Z)$ under the original measure can be evaluated under $P_{(\theta)}$ if we first multiply $g(Z)$ by the likelihood ratio $e^{-\theta Z + \psi_Z(\theta)}$, i.e.,

$$\mathsf{E}[g(Z)] = \mathsf{E}_{(\theta)}[g(Z) \cdot e^{-\theta Z + \psi_Z(\theta)}], \tag{21}$$

where $\mathsf{E}_{(\theta)}$ denotes expectation under $P_{(\theta)}$. From this, we get the following relation, which will be used very often in our proofs:

$$\mathsf{E}_{(\theta)}[Z] = \mathsf{E}[Ze^{\theta Z - \psi_Z(\theta)}] = \psi'_Z(\theta).$$

A twist can be applied to a sequence of i.i.d. random variables $\{Z_n\}$. Let $\tilde{P}$ be the probability measure obtained by θ-twisting $Z_1, Z_2, \ldots$. Then for any fixed n and any function g of $\{Z_1, \ldots, Z_n\}$, (21) generalizes to

$$\mathsf{E}[g(Z_1, \ldots, Z_n)] = \tilde{\mathsf{E}}\left[g(Z_1, \ldots, Z_n) \prod_{j=1}^{n} e^{-\theta Z_j + \psi_Z(\theta)} \right],$$

where $\tilde{\mathsf{E}}$ denotes expectation under $\tilde{P}$. This identity extends to stopping times—i.e., for any stopping time T and any function g of $\{Z_1, \ldots, Z_T\}$, *Wald's likelihood ratio identity* (see Siegmund 1985, p.166 or Asmussen 1987, p.258) gives

$$\mathsf{E}[g(Z_1, \ldots, Z_T);\, T < \infty] = \tilde{\mathsf{E}}\left[g(Z_1, \ldots, Z_T) \prod_{j=1}^{T} e^{-\theta Z_j + \psi_Z(\theta)};\, T < \infty \right].$$

We use this frequently. (A semicolon inside an expectation indicates that the expectation is evaluated over the event following the semicolon.)

Proof of Theorem 1. Because of the representation in (20), the distribution of $R(s)$ can be analyzed through W. To that end, let $X_n = \sum_{j=1}^{D_n} B_{n,j} - A_n$, for all $n \geq 1$; $S_0 = 0$ and $S_n = \sum_{j=1}^{n} X_j$. A classical result (see, e.g., Asmussen 1987, p. 80) states that W has the same distribution as $\max_{n \geq 0} S_n$. Thus,

$$P(R(s) > x) = P\left(W > \sum_{j=1}^{N(s)-1} \bar{A}_j - \sum_{j=1}^{H(s)} \bar{B}_j + x \right) = P\left(\max_{n \geq 0} S_n > \sum_{j=1}^{N(s)-1} \bar{A}_j - \sum_{j=1}^{H(s)} \bar{B}_{N(s),j} + x\right).$$

If we set $L = \sum_{j=1}^{N(s)-1} \bar{A}_j - \sum_{j=1}^{H(s)} \bar{B}_{N(s),j} + x$ and $T = \inf\{n \geq 1 : S_n > L\}$, then

$$P(R(s) > x) = P(T < \infty). \tag{22}$$

Let $\tilde{P}$ be the measure obtained by γ-twisting $X_1, X_2, \ldots$, β-twisting $\bar{D}_1, \bar{D}_2, \ldots$, $(-\gamma)$-twisting $\bar{A}_1, \bar{A}_2, \ldots$, and γ-twisting $\bar{B}_{N,1}, \bar{B}_{N,2}, \ldots$. Notice that $N(s)$ is a stopping time for $\{\bar{D}_j\}$ and T is a stopping time for $\{X_n\}$. In this setting, Wald's identity gives

$$\begin{aligned} P(T < \infty) &= \tilde{\mathsf{E}}\Bigg[\prod_{j=1}^{T} e^{-\gamma X_j + \psi_X(\gamma)} \prod_{j=1}^{N} e^{-\beta \bar{D}_j + \psi_D(\beta)} \\ &\qquad \cdot \prod_{j=1}^{N-1} e^{\gamma \bar{A}_j + \psi_A(-\gamma)} \prod_{j=1}^{H} e^{-\gamma \bar{B}_{N,j} + \psi_B(\gamma)};\, T < \infty \Bigg] \\ &= \tilde{\mathsf{E}}\Bigg[\exp\Bigg\{ -\gamma S_T - \beta \sum_{j=1}^{N} \bar{D}_j + N\psi_D(\beta) + \gamma \sum_{j=1}^{N-1} \bar{A}_j \\ &\qquad + (N-1)\psi_A(-\gamma) - \gamma \sum_{j=1}^{H} \bar{B}_{N,j} + H\beta \Bigg\};\, T < \infty \Bigg] \\ &= e^{-\gamma x - \beta s + \psi_D(\beta)} \tilde{\mathsf{E}}[e^{-\gamma(S_T - L)};\, T < \infty]. \end{aligned} \tag{23}$$

Under $\tilde{P}$, $\tilde{\mathsf{E}}[X_1] = \mathsf{E}[e^{\gamma X_1} X_1] = \psi'_X(\gamma) > 0$, so $\tilde{P}(T < \infty) = 1$. The random level L is independent of $\{S_n\}$ and $L \to \infty$ as $s + x \to \infty$. A minor extension of a classical result in renewal theory (see Corollary 8.33 of Siegmund 1985, or Theorem XII.5.2 of Asmussen 1987) shows that

$$C_1 = \lim_{s+x \to \infty} \tilde{\mathsf{E}}[e^{-\gamma(S_T - L)}] = \tilde{\mathsf{E}}[e^{-\gamma Z_e}]$$

exists, where Z_e has the equilibrium distribution of the ascending ladder heights of the random walk under $\tilde{P}$. (See Chapter 12 of Feller 1972, Chapter 1 of Prabhu 1980, Chapter VII of Asmussen 1987, or Chapter VIII of Siegmund 1985 for background on ladder heights and related results from the theory of random walks.) So we have

$$\lim_{s+x \to \infty} e^{\gamma x + \beta s} P(R(s) > x) = e^{\psi_D(\beta)} C_1 = C. \quad \square \tag{24}$$

When we have an assembly time U_n, the steady-state response time becomes $\hat{R}(s) = R(s) + U$, where $R(s)$ is as in Theorem 1. By the simple relation $P(\hat{R}(s) > x) = P(R(s) > x - U)$ and conditioning on U, the argument in the proof of Theorem 1 shows that

$$\lim_{s+x \to \infty} e^{\gamma x + \beta s} P(\hat{R}(s) > x | U) = \lim_{s+x \to \infty} e^{\gamma x + \beta s} P(R(s) > x - U | U) = Ce^{\gamma U}.$$

When $s \to \infty$ but x stays finite in the limit above, we require x to be such that $P(U \leq x) = 1$. Invoking the dominated convergence theorem, we get

$$\lim_{s+x \to \infty} e^{\gamma x + \beta s} P(\hat{R}(s) > x) = C\mathsf{E}[e^{\gamma U}],$$

which means that the assembly time changes only the constant and does not alter the asymptotic trade-off between x and s. Obviously, this is also true when assembly times are added in multiple-item systems.

The change of measure introduced above is also useful in estimating the service level for given x and s through relation (22). When $x + s$ is large, L is also large. But for a stable production system the random walk $\{S_n\}$ has negative drift, so $\{T < \infty\}$ is a rare event and it becomes

increasingly rare as the service level increases. Straightforward simulation is not efficient, if possible at all. Working with the new measure $\tilde{P}$, we appeal to (23) and estimate $P(T < \infty)$ by averaging i.i.d. replications of

$$e^{-\gamma(S_T - L) + \psi_D(\beta) - \gamma x - \beta s}$$

generated under $\tilde{P}$.

Proof of Proposition 1. We adopt the notation of Siegmund (1985). Let $\tau_+ = \min\{n \geq 1 : S_n \geq 0\}$, $\tau_- = \min\{n \geq 1 : S_n \leq 0\}$. By Equation 8.48 of Siegmund (1985)

$$\tilde{\mathsf{E}}[e^{-\gamma Z_e}] = P(\tau_+ = \infty)\tilde{P}(\tau_- = \infty)(\gamma\psi_X'(\gamma))^{-1}. \tag{25}$$

The first factor $P(\tau_+ = \infty) = P(\max_{n \geq 0} S_n = 0) = P(W = 0) = 1 - \lambda\mathsf{E}[D]\mathsf{E}[B]$. To evaluate $\tilde{P}(\tau_- = \infty)$, let $\hat{X}_n = -X_n$ and $\hat{S}_n = -S_n$, for $n \geq 1$; $\hat{S}_0 = 0$.

$$\begin{aligned}\tilde{P}(\tau_- = \infty) &= \tilde{P}(\min_{n \geq 0} S_n = 0) \\ &= \tilde{P}(\max_{n \geq 0} \hat{S}_n = 0) \\ &= \tilde{P}(\hat{W} = 0),\end{aligned}$$

where $\hat{W}$ is the steady state waiting time of a $G/M/1$ queue with interarrival time $\sum_{j=1}^{D} B_j$ and service time A. Notice that $\mathsf{E}[X_1] > 0$, so the $G/M/1$ queue is in fact stable under $\tilde{P}$. By Theorem 1.3. of Asmussen (1987, p. 204)

$$\tilde{P}(\hat{W} = 0) = 1 - \tilde{\mathsf{E}}\left[\exp\left\{-\gamma\left(\sum_{j=1}^{D} B_j\right)\right\}\right] = 1 - \mathsf{E}[e^{-\gamma A}]$$

$$= \frac{\gamma}{\lambda + \gamma}.$$

Write $\psi_X'(\gamma) = \psi_D'(\psi_B(\gamma)) \cdot \psi_B'(\gamma) - \psi_A'(-\gamma) = \psi_D'(\beta)\psi_B'(\gamma) - (\lambda + \gamma)^{-1}$ and $e^{\psi_D(\beta)} = e^{-\psi_A(-\gamma)} = (\lambda + \gamma)/\lambda$. Recalling (24) and (25) we have

$$\begin{aligned}C &= e^{\psi_D(\beta)}(1 - \lambda\mathsf{E}[D]\mathsf{E}[B])\frac{\gamma}{\lambda + \gamma}\gamma^{-1} \\ &\quad \cdot \left(\psi_D'(\beta)\psi_B'(\gamma) - \frac{1}{\lambda + \gamma}\right)^{-1} \\ &= \lambda^{-1}(\lambda + \gamma)(1 - \lambda\mathsf{E}[D]\mathsf{E}[B]) \\ &\quad \cdot (\psi_D'(\beta)\psi_B'(\gamma)(\lambda + \gamma) - 1)^{-1}.\end{aligned}$$

□

4. MULTIPLE-ITEM SYSTEMS

We turn next to the setting of Theorem 2, in which the nth order requires D_n^i units of item i. The vectors $(D_n^1, \ldots, D_n^d)$, $n = 1, 2, \ldots$ are i.i.d., but their components need not be independent. An order is filled only when all units of all items required are available. Thus the response time of the nth order is given by $R_n = \max\{R_n^1, \ldots, R_n^d\}$. Under the stability conditions, $\mathsf{E}[B^i]\mathsf{E}[D^i] - \mathsf{E}[A] < 0$, for all $i = 1, \ldots, d$, $(R_n^1, \ldots, R_n^d)$ converges in distribution to a steady-state limit $(R^1, \ldots, R^d)$ as $n \to \infty$, and the steady-state order response time is $R = \max\{R^1, \ldots, R^d\}$. The tail probability of each R^i is described by Theorem 1. From this and the simple bounds

$$\max_i P(R^i(s) > x) \leq P(R(s) > x) \leq \sum_i P(R^i(s) > x), \tag{26}$$

it follows directly that

$$\lim_{x \to \infty} \frac{1}{x} \log P(R(s) > x) = -\gamma$$

and

$$\lim_{s \to \infty} \frac{1}{s} \log P(R(s) > x) = -\alpha,$$

with $\gamma = \min_i \gamma_i$ and $\alpha = \min_i \alpha_i$. Strengthening these logarithmic limits to the exponential asymptotics claimed in Theorem 2 requires a more involved argument. We shall need the following auxiliary result for the bivariate cumulant generating function of random vector (Y_1, Y_2), defined by

$$\psi_{Y_1,Y_2}(\theta_1, \theta_2) = \log \mathsf{E}[e^{\theta_1 Y_1 + \theta_2 Y_2}].$$

We use $\partial_{\theta_1}\psi$, $\partial_{\theta_2}\psi$ to denote the partial derivatives of ψ with respect to its first and second argument, respectively.

Lemma 2. *Suppose for some $\omega_1 > 0$ and $\omega_2 > 0$, $\psi_{Y_1,Y_2}(\omega_1, 0) = \psi_{Y_1,Y_2}(0, \omega_2) = c < \infty$ and $(\omega_1, 0)$ is in the interior of the domain of ψ_{Y_1,Y_2}, then*

$$\omega_1 \partial_{\theta_1}\psi_{Y_1,Y_2}(\omega_1, 0) \geq \omega_2 \partial_{\theta_2}\psi_{Y_1,Y_2}(\omega_1, 0), \tag{27}$$

and equality holds if and only if $\omega_1 Y_1 = \omega_2 Y_2$, a.s.

Proof. Let $\mathcal{A} = \{(\theta_1, \theta_2) : \psi_{Y_1,Y_2}(\theta_1, \theta_2) \leq c\}$. Convexity of ψ_{Y_1,Y_2} implies convexity of $\mathcal{A}$ which further implies (essentially by the supporting hyperplane theorem—see the argument used for Theorem 2.3.7 of Bazaraa and Shetty 1979) that for any $(u, v) \in \mathcal{A}$

$$\begin{aligned}&(u - \omega_1)\partial_{\theta_1}\psi_{Y_1,Y_2}(\omega_1, 0) + (v - 0)\partial_{\theta_2}\psi_{Y_1,Y_2}(\omega_1, 0) \\ &\quad \leq 0.\end{aligned}$$

Applying this inequality at $(0, \omega_2) \in \mathcal{A}$ we get (27).

When the equality holds, $\psi_{Y_1,Y_2}(\theta_1, \theta_2) = c$, for all (θ_1, θ_2) on the line segment $[(0, \omega_2), (\omega_1, 0)]$, i.e.,

$$\psi_{Y_1,Y_2}\left(\omega_1 - \frac{\omega_1}{\omega_2}\theta_2, \theta_2\right) = c, \quad \forall\, 0 \leq \theta_2 \leq \omega_2,$$

$$\mathsf{E}[e^{\omega_1 Y_1 - \psi_{Y_1,Y_2}(\omega_1,0)} \cdot e^{(Y_2 - \omega_1/\omega_2 Y_1)\theta_2}] = 1, \quad \forall\, 0 \leq \theta_2 \leq \omega_2. \tag{28}$$

Notice that $\mathsf{E}[e^{\omega_1 Y_1 - \psi_{Y_1,Y_2}(\omega_1,0)}] = 1$, so $e^{\omega_1 Y_1 - \psi_{Y_1,Y_2}(\omega_1,0)}$ defines the likelihood ratio between two probability measures P and $\tilde{P}_{(\omega_1,0)}$, which are absolutely continuous with respect to each other. Equation (28) can be written as $\tilde{\mathsf{E}}_{(\omega_1,0)}[e^{(Y_2 - (\omega_1/\omega_2)Y_1)\theta_2}] = 1$ for all $0 \leq \theta_2 \leq \omega_2$. As a result, we have $\omega_2 Y_2 - \omega_1 Y_1 = 0$, $\tilde{P}_{(\omega_1,0)}$-a.s., and P-a.s. □

We can now give the proof of Theorem 2.

Proof of Theorem 2. The limits on x and s follow fairly directly from Theorem 1 and simple bounds if $|\mathcal{I}_x| = 1$ or $|\mathcal{I}_s| = 1$, respectively. Most of the difficulty arises from dealing with ties among constraining items.

We first prove (6), the limit on x. By the simple upper bound in (26), we have $\limsup_{x\to\infty} e^{\gamma x}P(R(s)>x) \le \sum_{i\in\mathcal{J}_x} C_i e^{-\alpha_i s}$. For the lower bound, we refine (26) to

$$P(R(s)>x) \ge \sum_i P(R^i>x) - \sum_{i\ne j} P(R^i(s)>x,\, R^j(s)>x),$$

and get

$$\liminf_{x\to\infty} e^{\gamma x}P(R(s)>x) \ge \sum_{i\in\mathcal{J}_x} C_i e^{-\alpha_i s} - \sum_{i,j\in\mathcal{J}_x, i\ne j} \limsup_{x\to\infty} e^{\gamma x}P(R^i(s)>x,\, R^j(s)>x).$$

From these bounds, (6) is obviously true if $|\mathcal{J}_x| = 1$. We consider the case $|\mathcal{J}_x| \ge 2$ and index two arbitary items in $\mathcal{J}_x$ by 1 and 2, i.e., $\gamma_1 = \gamma_2 = \gamma$. It suffices to show

$$\limsup_{x\to\infty} e^{\gamma x}P(R^1(s)>x,\, R^2(s)>x) = 0.$$

Indeed, it suffices to show

$$\limsup_{x\to\infty} e^{\gamma x}P(R^1(s)>x,\, R^2(s)>x,\, N_2\le N_1) = 0, \quad (29)$$

where the N_i are defined for each item i paralleling (18): $N_i = \min\{n\ge 1 : \sum_{j=1}^n \bar{D}_j^i > s^i\}$. We continue to use the notation in the proof of Theorem 1 on each item and analyze the response times through queues and random walks, which are now *correlated* across items.

We represent the response time distribution of item 1 as

$$R^1 \overset{\mathcal{D}}{=} \left((\max_{n\ge 0} S_n^1) - \sum_{j=1}^{N_1-1} \bar{A}_j + \sum_{j=1}^{H_1} \bar{B}^1_{N_1,j} \right)^+,$$

where $S_0^1 = 0$, $S_n^1 = \sum_{j=1}^n X_j^1$, for $n \ge 1$ with

$$X_n^1 = \sum_{j=1}^{D_n^1} B^1_{n,j} - A_n. \quad (30)$$

Then, on the set $\{N_2 \le N_1\}$, R^2 has the representation

$$R^2 \overset{\mathcal{D}}{=} \left((\max_{n\ge 0} S_n^2) - \sum_{j=1}^{N_2-1} \bar{A}_j + \sum_{j=1}^{H_2} \bar{B}^2_{N_2,j} \right)^+,$$

where $S_0^2 = 0$, $S_n^2 = \sum_{j=1}^n X_j^2$, for $n \ge 1$ with

$$X_n^2 = \begin{cases} \sum_{j=1}^{D^2_{n+N_2}} \bar{B}^2_{n+N_2,j} - \bar{A}_{n+N_2}, \\ \quad \text{for } 1 \le n \le N_1 - N_2, \\ \sum_{j=1}^{D^2_{n-(N_1-N_2)}} B^2_{n-(N_1-N_2),j} - A_{n-(N_1-N_2)}, \\ \quad \text{for } n > N_1 - N_2. \end{cases} \quad (31)$$

To see why R^1, R^2 have these representations, notice that R^1, R^2 are the steady-state response times for item 1 and item 2 of the *same* order. As a result, X_n^2 is correlated with $X^1_{n-(N_1-N_2)}$ for $n > N_1 - N_2$. We define $L_i = \sum_{j=1}^{N_i-1} \bar{A}_j - \sum_{j=1}^{H_i} \bar{B}^i_{N_i} j + x$ and $T_i = \inf\{n \ge 1 : S_n^i > L_i\}$, $i = 1, 2$.

To analyze the left side of (29), we let $\tilde{P}^1_{T_1}$ be the measure defined by γ-twisting $\{X_1^1, \ldots, X_{T_1}^1\}$, β_1-twisting $\{\bar{D}_1^1, \bar{D}_2^1, \ldots\}$, $(-\gamma)$-twisting $\{\bar{A}_1, \bar{A}_2, \ldots\}$, and γ-twisting $\{\bar{B}^1_{N_1,1}, \ldots, \bar{B}^1_{N_1,H_1}\}$. Then for any event E determined by $\{X_1^1, \ldots, X_{T_1}^1;\ \bar{D}_1^1, \ldots, \bar{D}_{N_1}^1;\ \bar{A}_1, \ldots, \bar{A}_{N_1-1};\ \bar{B}^1_{N_1,1}, \ldots, \bar{B}^1_{N_1,H_1}\}$,

$$\begin{aligned} P(E) &= \tilde{\mathsf{E}}^1_{T_1}\Big[\exp\Big\{ -\gamma S^1_{T_1} - \beta_1 \sum_{j=1}^{N_1} D_j^1 + N_1 \psi_{D^1}(\beta_1) \\ &\qquad + \gamma \sum_{j=1}^{N_1-1} \bar{A}_j + (N_1-1)\psi_A(-\gamma) \\ &\qquad - \beta_1 \sum_{j=1}^{H_1} B^1_{N_1,j} + H_1 \psi_{B^1}(\gamma) \Big\};\, E \Big] \\ &= \tilde{\mathsf{E}}^1_{T_1}[e^{-\gamma(S^1_{T_1}-L_1)};\, E] \cdot e^{-\gamma x - \beta_1 s^1 + \psi_{D^1}(\beta_1)}. \end{aligned}$$

Both N_1 and N_2 are finite a.s. To show (29), we write

$$\begin{aligned} &e^{\gamma x}P(R^1(s)>x,\, R^2(s)>x,\, N_2\le N_1) \\ &= e^{\gamma x}P(T_1<\infty,\, T_2<\infty,\, N_2\le N_1) \\ &= e^{\gamma x}\tilde{\mathsf{E}}^1_{T_1}[e^{-\gamma_1(S^1_{T_1}-L_1)};\, T_1<\infty,\, T_2<\infty,\, N_2\le N_1] \\ &\quad \cdot e^{-\gamma x - \alpha_1 s + \psi_{D^1}(\beta_1)} \\ &\le \tilde{P}_{T_1}(T_1<\infty,\, T_2<\infty,\, N_2\le N_1) e^{-\alpha_1 s + \psi_{D^1}(\beta_1)} \\ &= \tilde{P}_{T_1}(T_2 \le T_1 + (N_1-N_2) < \infty,\, N_2\le N_1) \\ &\quad \cdot e^{-\alpha_1 s + \psi_{D^1}(\beta_1)} + \tilde{P}_{T_1}(T_1 + (N_1-N_2) < T_2 < \infty, \\ &\quad N_2 \le N_1) e^{-\alpha_1 s + \psi_{D^1}(\beta_1)}. \end{aligned} \quad (32)$$

We will show that both terms in (32) have limit 0 when $x \to \infty$. This is obviously true if $\tilde{\mathsf{E}}^1_{T_1}[X^2] < 0$ because $\tilde{P}_{T_1}(T_2<\infty) \to 0$. If $\tilde{\mathsf{E}}^1_{T_1}[X^2] \ge 0$, then

$$\begin{aligned} &\frac{T_2}{T_1 + (N_1-N_2)} \\ &= \frac{1}{1 + (N_1-N_2)/T_1} \cdot \frac{L_1/T_1}{L_2/T_2} \cdot \frac{L_2}{L_1} \overset{\text{a.s.}}{\to} 1 \cdot \frac{\tilde{\mathsf{E}}^1_{T_1}[X^1]}{\tilde{\mathsf{E}}^1_{T_1}[X^2]} \cdot 1 \\ &= \frac{\mathsf{E}[X^1 e^{\gamma_1 X^1}]}{\mathsf{E}[X^2 e^{\gamma_1 X^1}]} = \frac{\partial_{\theta_1}\psi_{X^1,X^2}(\gamma_1, 0)}{\partial_{\theta_2}\psi_{X^1,X^2}(\gamma_1, 0)} > \frac{\gamma_2}{\gamma_1} = 1, \end{aligned}$$

where the last inequality uses Lemma 2. The limiting ratio above equals ∞ when $\tilde{\mathsf{E}}^1_{T_1}[X^2] = 0$. We conclude that the first term of (32) goes to 0.

What remains to be shown is that the second term of (32) also goes to 0. Observe that on the set $\{T_1 + (N_1 - N_2) < T_2\}$, the process $\{S_n^2 - S^2_{T_1+(N_1-N_2)},\, n > T_1 + (N_1 - N_2)\}$ has the original law under $\tilde{P}_{T_1}$ and thus has negative drift (cf. (31) to see how $\{S_n^2\}$ is generated). It is independent of $\{S_n^2,\, n \le T_1 + (N_1 - N_2)\}$ hence independent of $L_2 - S^2_{T_1+(N_1-N_2)}$. Since

$$\begin{aligned} &\tilde{P}^1_{T_1}(T_1 + (N_1-N_2) < T_2 < \infty,\, N_2 \le N_1) \\ &\le \tilde{P}^1_{T_1}\Big(\max_{n > T_1+(N_1-N_2)} S_n^2 > L_2,\, N_2 \le N_1 \Big) \\ &= \tilde{P}^1_{T_1}\Big(\max_{n > T_1+(N_1-N_2)} (S_n^2 - S^2_{T_1+(N_1-N_2)}) > L_2 \\ &\qquad - S^2_{T_1+(N_1-N_2)},\, N_2 \le N_1 \Big), \end{aligned}$$

the result follows if $L_2 - S^2_{T_1}+(N_1-N_2) \to \infty$. This is indeed the case, because

$$\frac{L_2}{S^2_{T_1+(N_1-N_2)}} = \frac{L_1/T_1}{S^2_{T_1+(N_1-N_2)}/(T_1+N_1-N_2)} \cdot \frac{1}{1+(N_1-N_2)/T_1} \cdot \frac{L_2}{L_1} \overset{\text{a.s.}}{\to} \frac{\tilde{\mathsf{E}}^1_{T_1}[X^1]}{\tilde{\mathsf{E}}^1_{T_1}[X^2]} \cdot 1 \cdot 1 > 1.$$

the limiting ratio above equals ∞ when $\tilde{\mathsf{E}}^1_{T_1}[X^2] = 0$. This completes the proof of (6).

The limit on s in (7) is a bit more involved due to the way random levels L_i go to ∞ as $s \to \infty$. We give an outline of the proof here and refer the reader to Wang (1997) for the complete proof. Similar to the proof of (6), we only need to consider the case $|\mathcal{I}_s| \geq 2$ where we index two arbitrary items in $\mathcal{I}_s$ by 1 and 2, i.e., $\alpha_1 \triangleq k_1\beta_1 = \alpha_2 \triangleq k_2\beta_2 = \alpha$. It then suffices to show

$$\limsup_{s\to\infty} e^{\alpha s} P(T_1 < \infty,\ T_2 < \infty,\ N_2 \leq N_1) = 0. \tag{33}$$

We consider the two cases $\gamma_1 > \gamma_2$ and $\gamma_1 \leq \gamma_2$ separately and show that (33) always holds. If $\gamma_1 > \gamma_2$, we write the left side of (33) as

$$\begin{aligned} & e^{\alpha s} P(T_1 < \infty,\ T_2 < \infty,\ N_2 \leq N_1) \\ & \leq e^{\alpha s} P(T_1 < \infty,\ N_2 \leq N_1) \\ & = e^{\alpha s} \tilde{\mathsf{E}}^1_{T_1}[e^{-\gamma_1(S^1_{T_1}-L_1)};\ T_1 < \infty,\ N_2 \leq N_1] e^{-\gamma_1 x - \alpha s + \psi_{D^1}(\beta_1)} \\ & \leq e^{-\gamma_1 x + \psi_{D^1}(\beta_1)} \tilde{P}^1_{T_1}(N_2 \leq N_1). \end{aligned}$$

Under $\tilde{P}^1_{T_1}$, we use the strong law of large numbers to show that as $s \to \infty$, N_2/N_1 has a limit larger than 1. So $\tilde{P}^1_{T_1}(N_2 \leq N_1) \to 0$ and (33) follows when $\gamma_1 > \gamma_2$.

For the case $\gamma_1 \leq \gamma_2$, which implies $\eta_1 = -\psi_A(-\gamma_1) \leq \eta_2 = -\psi_A(-\gamma_2)$, we analyze the left side of (33) through a different change of measure. Let $\tilde{P}^2_{T_2,N_2}$ be the measure obtained by γ_2-twisting $\{X^2_1, \ldots, X^2_{T_2}\}$, β_2-twisting $\{\bar{D}^2_1, \ldots, \bar{D}^2_{N_2}\}$, $(-\gamma_2)$-twisting $\{\bar{A}_1, \ldots, \bar{A}_{N_2-1}\}$, and γ_2-twisting $\{\bar{B}^2_{N_1}, 1, \ldots, \bar{B}^2_{N_2}, H_2\}$. We write the left side of (33) as

$$\begin{aligned} & e^{\alpha s} P(T_1 < \infty,\ T_2 < \infty,\ N_2 \leq N_1) \\ & = e^{\alpha s} \tilde{\mathsf{E}}^2_{T_2,N_2}[e^{-\gamma_2(S^2_{T_2}-L_2)};\ T_1 < \infty,\ T_2 < \infty,\ N_2 \leq N_1] \\ & \quad \cdot e^{-\gamma_2 x - \alpha s + \psi_{D^2}(\beta_2)} \\ & \leq \tilde{P}^2_{T_2,N_2}(T_1 < \infty,\ T_2 < \infty,\ N_2 \leq N_1) e^{-\gamma_2 x + \psi_{D^2}(\beta_2)} \\ & = e^{-\gamma_2 x + \psi_{D^2}(\beta_2)} \tilde{P}^2_{T_2,N_2}(T_2 \leq N_1 - N_2,\ T_1 < \infty,\ N_2 \leq N_1) \\ & \quad + e^{-\gamma_2 x + \psi_{D^2}(\beta_2)} \tilde{P}^2_{T_2,N_2}(T_1 + (N_1 - N_2) \leq T_2 < \infty,\ N_2 \leq N_1) \\ & \quad + e^{-\gamma_2 x + \psi_{D^2}(\beta_2)} \tilde{P}^2_{T_2,N_2}(0 < T_2 - (N_1 - N_2) < T_1 < \infty,\ N_2 \leq N_1), \end{aligned} \tag{34}$$

and show that each of the three terms in (34) has a limit of 0. (It suffices to consider the case $\tilde{\mathsf{E}}_{T_2,N_2}[X^1] \geq 0$ because otherwise $\tilde{P}_{T_2,N_2}(T_1 < \infty) \to 0$ trivially.) The first term is easy to show because on the set $\{T_2 < N_1 - N_2\}$ the process $\{S^1_n;\ n \geq 1\}$ has the original law under $\tilde{P}^2_{T_2,N_2}$, thus has negative drift. The random level $L_1 \to \infty$, $\tilde{P}_{T_2,N_2}$-a.s., so $\tilde{P}_{T_2,N_2}(T_1 < \infty) \to 0$.

The arguments for the second and third terms of (34) parallel those for the first and second terms of (32), respectively. For the second term, we use the strong law of large numbers to show that under $\tilde{P}_{T_2,N_2}$, $(T_1 + N_1 - N_2)/T_2$ has a limit that is larger than 1 under the condition

$$\frac{\mathsf{E}[D^2 e^{\beta_2 D^2 - \psi_{D^2}(\beta_2)}]}{\mathsf{E}[D^1 e^{\beta_2 D^2 - \psi_{D^2}(\beta_2)}]} \geq \frac{\beta_1}{\beta_2^-}, \tag{35}$$

where β_2^- $(\leq \beta_2)$ is defined by $\psi_{D^2}(\beta_2^-) = \eta_1$ $(\leq \eta_2)$. (The existence of β_2^- is guaranteed by our standing assumption that $\psi_{D^2}(\theta)$ does not jump to ∞ as θ increases.) As a result, we have $\tilde{P}_{T_2,N_2}(T_1 + (N_1 - N_2) \leq T_2,\ N_2 \leq N_1) \to 0$. The condition in (35) is implied by the independence of D^i. To see this, notice that the independence of D^i implies $\mathsf{E}[D^1 e^{\beta_2 D^2 - \psi_{D^2}(\beta_2)}] = \mathsf{E}[D^1] = \psi'_{D^1}(0) \leq \eta_1/\beta_1$; convexity of ψ implies $\mathsf{E}[D^2 e^{\beta_2 D^2 - \psi_{D^2}(\beta_2)}] = \psi'_{D^2}(\beta_2) \geq \psi'_{D^2}(\beta_2^-) \geq \eta_1/\beta_2^-$. Hence (35) follows.

For the third term of (34), observe that on the set $(0 < T_2 - (N_1 - N_2) < T_1)$, the process $\{S^1_n - S^1_{T_2} - (N_1 - N_2),\ n > T_2 - (N_1 - N_2)\}$ has the original law under $\tilde{P}^2_{T_2,N_2}$ and thus has negative drift. Also notice that

$$\begin{aligned} & \tilde{P}^2_{T_2,N_2}(0 < T_2 - (N_1 - N_2) < T_1 < \infty,\ N_2 \leq N_1) \\ & \leq \tilde{P}^2_{T_2,N_2}\Big(\max_{n > T_2 - (N_1 - N_2)} S^1_n > L_1,\ N_2 \leq N_1\Big) \\ & = \tilde{P}^2_{T_2,N_2}\Big(\max_{n > T_2 - (N_1 - N_2)} (S^1_n - S^1_{T_2 - (N_1 - N_2)}) \\ & \qquad > L_1 - S^1_{T_2 - (N_1 - N_2)},\ N_2 \leq N_1\Big). \end{aligned}$$

This probability goes to 0 if $L_1 - S^1_{T_2} - (N_1 - N_2) \to \infty$, and this condition holds because we can show that $L_1/S^1_{T_2} - (N_1 - N_2)$ has a limit that is larger than 1 under condition (35). □

It is worth noting that independence of the D_i was used only to verify the rather technical condition in (35) and does not appear to be essential to the result itself. We have not, however, found a conveniently stated weaker condition to cover the case $|\mathcal{I}_s| > 1$ and replace independence.

For Theorem 3, the setting reduces to that of Theorem 2 when we consider product j and the set $\mathcal{I}_j$ of items it requires. The results follow by the same argument used in the proof of Theorem 2.

5. CONCLUDING REMARKS

We have demonstrated both theoretically and numerically that it is possible to quantify the trade-off between longer leadtimes or higher inventory levels in achieving a target fill rate, in a class of production-inventory models. Not

surprisingly, the trade-off is sharpest in single-item systems. When multiple items are assembled into multiple products, the trade-off depends in part on which items most constrain the product-level fill rate. One aspect of our analysis is a characterization of which items are most constraining at higher inventory levels and which are most constraining at longer leadtimes. This distinction potentially offers a new perspective on where efforts should be expended to improve service; it emerges naturally from an analysis focused directly on service levels.

ACKNOWLEDGMENT

This work is supported by NSF grant DMI-94-57189.

REFERENCES

ASMUSSEN, S. 1987. *Applied Probability and Queues*. Wiley, New York.

BAKER, K. R., M. J. MAGAZINE, AND H. NUTTLE. 1986. The Effect of Commonality on Safety Stock in a Simple Inventory Model. *Mgmt. Sci.* **32,** 982–988.

BAZARAA, M. S. AND C. M. SHETTY. 1979. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, New York.

BOWERSOX, D. J. AND M. B. COOPER. 1992. *Strategic Marketing Channel Management*. McGraw-Hill, New York.

BUZACOTT, J. A. AND J. G. SHANTHIKUMAR. 1994. Safety Stock versus Safety Time in MRP Controlled Production Systems. *Mgmt. Sci.* **40,** 1678–1689.

CHANG, C. A. 1985. The Interchangeability of Safety Stocks and Safety Time. *J. Oper. Mgmt.* **6,** 35–42.

CHASE, R. B. AND N. J. AQUILANO. 1995. *Production and Operations Management: Manufacturing and Service*, Seventh Edition. Irwin, Chicago.

CLARK, A. J., AND H. SCARF. 1960. Optimal Policies for a Multi-Echelon Inventory. *Mgmt. Sci.* **6,** 475–490.

ETTL, M., G. E. FEIGIN, G. Y. LIN, AND D. D. YAO. 1995. A Supply-Chain Model with Base-Stock Control and Service Level Requirements. IBM Research Report, IBM T. J. Watson Research Center, Yorktown Heights, New York.

FEDERGRUEN, A. AND P. ZIPKIN. 1986a. An Inventory Model with Limited Production Capacity and Uncertain Demands, I: The Average Cost Criterion. *Math. O. R.* **11,** 193–207.

FEDERGRUEN, A. AND P. ZIPKIN. 1986b. An Inventory Model with Limited Production Capacity and Uncertain Demands, II: The Discounted Cost Criterion. *Math. O. R.* **11,** 208–215.

FELLER, W. 1971. *An Introduction to Probability Theory and its Applications*, Vol. 2, Second Edition. Wiley, New York.

GLASSERMAN, P. 1997. Bounds and Asymptotics for Planning Critical Safety Stocks. *Oper. Res.* **45,** 244–257.

GUT, A. 1988. *Stopped Random Walks*. Springer, New York.

HAUSMAN, W. H., H. L. LEE, AND A. X. ZHANG. 1993. Order Response Time Reliability in a Multi-Item Inventory System. Working Paper, Department of Industrial Engineering and Engineering Management, Stanford University, Stanford, CA.

HESKETT, J. L., W. E. SASSER, JR., AND C. W. L. HART. 1990. *Service Breakthroughs: Changing the Rules of the Game*. The Free Press, New York.

KENDALL, M. 1987. *Advanced Theory of Statistics*, Vol. II, Fifth Edition. Oxford, New York.

LEE, Y. AND P. H. ZIPKIN. 1992. Tandem Queues with Planned Inventories. *Oper. Res.* **40,** 936–947.

LIU, T. W. 1995. Analysis and Simulation of a Multistage Production-Inventory System. Ph.D. Thesis, Graduate School of Business, Columbia University, New York.

MCLAIN, J. O., L. J. THOMAS, AND J. B. MAZZOLA. 1992. *Operations Management: Production of Goods and Services*. Third Edition. Prentice-Hall, Englewood Cliffs, NJ.

MUCKSTADT, J. A. AND L. J. THOMAS. 1980. Are Multi-Echelon Inventory Methods Worth Implementing in Systems with Low-Demand Rate Items? *Mgmt. Sci.* **26,** 483–494.

MUCKSTADT, J. A. AND L. J. THOMAS. 1983. Improving Inventory Productivity in Multilevel Distribution Systems, in *Productivity and Efficiency in Distribution Systems*. D. Gautschi, Ed., Esleview North-Holland, New York.

PRABHU, N. U. 1980. *Stochastic Storage Processes: Queues, Insurance Risk, and Dams*. Springer, New York.

ROSLING, K. 1989. Optimal Inventory Policies for Assembly Systems Under Random Demands. *Oper. Res.* **37,** 565–579.

SCHRANER, E. 1996. Capacity/Inventory Trade-Offs in Assemble-to-Order Systems. Working Paper, IBM T.J. Watson Research Center, Yorktown Heights, New York.

SIEGMUND, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York.

SONG, J. S. 1997. On the Order Fill Rate in a Multi-Item Inventory System. Working Paper, Department of Industrial Engineering and Operations Research, Columbia University, New York, 1997. To appear in *Oper. Res*.

SONG, J. S., S. H. XU, AND B. LIU. 1996. Order-Fulfillment Performance Measures in an Assemble-to-Order System with Stochastic Leadtime. Working Paper, Department of Industrial Engineering and Operations Research, Columbia University, New York, 1997. To appear in *Oper. Res*.

VAN HOUTUM, G. J., K. INDERFURTH, AND W. H. M. ZIJM. 1995. Materials Coordination in Stochastic Multi-Echelon Systems. Working Paper LPOM-95-15, Department of Mechanical Engineering, University of Twente, The Netherlands. To appear in *Euro. J. Oper. Res*.

VEATCH, M. AND L. WEIN. 1994. Optimal Control of a Two-Station Tandem Production-Inventory System. *Oper. Res.* **42,** 337–350.

TAYUR, S. 1993. Computing the Optimal Policy for Capacitated Inventory Models. *Stochastic Models* **9,** 585–598.

WANG, Y. 1998. Service Levels in Assemble-to-Order Systems. Ph.D. thesis, Graduate School of Business. Columbia University, New York.

ZHANG, W. 1996. Production-Inventory Networks with Constant Processing Times. Ph.D. Thesis, Graduate School of Business, Columbia University, New York.

ZIPKIN, P. H. 1986. Models for Design and Control of Stochastic Multi-Item Batch Production Systems. *Oper. Res.* **34,** 91–104.

ZIPKIN, P. H. 1991. Does Manufacturing Need a JIT Revolution? *Harvard Business Review*, Jan.–Feb., 91111.