BOUNDS AND ASYMPTOTICS FOR PLANNING CRITICAL SAFETY STOCKS

PAUL GLASSERMAN

Columbia Business School, New York, New York

(Received April 1993; revisions received October 1993, June 1994; accepted April 1995)

We develop bounds and approximations for setting base-stock levels in production-inventory systems with limited production capacity. Our approximations become exact as inventories become *critical*, meaning either that the target service level is very high or the backorder penalty is very large. Our bounds apply even without this requirement. We consider both single-stage and multi-stage systems. For single-stage systems, we find tight bounds and asymptotically exact approximations for optimal base-stock levels; for multistage systems, our results give partial characterizations of the optimal levels. Part of our analysis is a precise connection, in the critical regime, between a multistage system and an associated single-stage system consisting solely of the bottleneck facility.

One purpose of inventory is to ensure that certain types of events—stockouts, large backorders, lost sales—are rare. If, say, backorder penalties are very large or the target service level is very high, then such events must be very rare. In some cases, rarity facilitates the analysis of a model through simplifications that emerge when rare events become extremely rare.

We develop and exploit this principle in an analysis of single- and multistage production-inventory systems with limited production capacity. The constraint on capacity, which could also be interpreted as a limitation on order size, complicates the problem of setting safety stocks either to ensure a certain level of service or to minimize costs. We show that in single-stage systems the required safety stocks admit simple approximations that become exact as inventories become *critical*, meaning that a target service level becomes high or, equivalently, that a backorder penalty becomes large. We supplement these approximations with bounds that remain valid over a wide range of parameters, not just in the critical regime. For multistage systems, we establish analogous asymptotics and bounds; however, the step from these results to the required stock levels is not as direct as it is for single-stage systems.

Underlying all our approximations is a result stating that tail probabilities associated with shortages decrease exponentially fast as safety stocks increase. The rate of this exponential decrease depends on the distribution of demands and on the system capacity, but is easily evaluated. Inverting exponential approximations to probabilities of shortages results in logarithmic approximations to stock levels required to meet a service objective or minimize a cost function. These inverted approximations are also asymptotically exact. A simple modification of the approximations results in upper and lower bounds, differing only by a constant from the exact asymptotics. Examples indicate that the gap between these bounds is often small. The bounds appear to be most effective when utilization is not too low.

In more detail, the models we consider have the following features. A single type of item is produced either by a single facility or by several facilities in series. Inventories are reviewed at intervals of fixed length; demands within each period follow a fairly arbitrary distribution but are assumed independent from period to period. Demands not met from stock are backordered. After the total demand in a period is revealed, production is set to try to restore inventory to a specified target, called a *base-stock level*. However, production in a single period may not suffice in reaching the target, because of a capacity constraint. Additionally, in our multistage model each stage draws raw material from upstream stages; the possible depletion of upstream inventories further constrains production in each period.

For the single-stage version of this model, Federgruen and Zipkin (1986a, b) show that a base-stock policy is in fact optimal; Tayur (1993) discusses computation of the optimal base-stock level. Clark and Scarf (1960) establish the optimality of base-stock policies in serial, *uncapacitated* multistage systems; see Rosling (1989) for more general topologies. These results (and ease of implementation) make base-stock policies natural candidates for multistage capacitated systems. Veatch and Wein (1994) show experimentally that base-stock policies are often close to optimal in a class of two-stage capacitated models; see also Lee and Zipkin (1992). A simulation-based optimization procedure and related stability issues for the model considered here are investigated in Glasserman and Tayur (1994, 1995), which may be consulted for further references.

The principal tool for the analysis in this paper is a set of techniques developed, to some extent in parallel, in risk theory, queueing theory and sequential analysis. These techniques provide approximations to tail probabilities associated with random walks. Key sources include Asmussen (1987, Chapter XII), Feller (1971, Chapter XII) and Siegmund (1985); these texts include references to earlier

Subject classifications: Inventory/production: service level approximations. Probability applications: rare events.
Area of review: Stochastic Processes and Their Applications.

0030-364X/97/4502-0244 $05.00

work in corresponding application areas. In our single-stage system, the stationary *shortfall*—the amount by which the base-stock level exceeds net inventory—has the distribution of the maximum of a random walk with negative drift, as does the stationary waiting time in a single-server queue. In our multistage system, the shortfall for each echelon has the distribution of the maximum over several dependent random walks. Through this link we obtain asymptotically exact approximations to tail probabilities for shortfalls, and hence to required base stocks. Related techniques have recently been used to analyze overflow probabilities in telecommunications systems; see Chang (1994), Whitt (1993), and references there.

This paper is organized to make the most important results as immediate as possible. A reader interested in the practical consequences of our analysis will find them in the first three sections; theoretical developments and longer proofs are postponed to the end. The single-stage system is treated in Section 1, beginning with a detailed formulation of the model, proceeding with the asymptotics and bounds, and concluding with extensions to variable production capacity. Section 2 treats the computation of bounds in more detail and includes numerical examples illustrating their performance. Section 3 extends the results of Section 1 to multistage systems. Background on random walks and proofs of our main results appear in Section 4.

1. THE SINGLE-STAGE SYSTEM

1.1. Shortfall Formulation

We consider a storage facility supplying external demands and receiving stock from a production facility. Time is divided into periods of fixed length. In each period, demands arrive and are either filled or backordered. The system operates under a base-stock policy in which production is set in each period to restore inventory to a target level s while not exceeding the per-period capacity c of the production facility. Thus, if I_n denotes the net inventory (on-hand inventory minus backorders) at the start of period n, and if D_n is the demand in period n, then production in period n is $\min\{c, s - I_n + D_n\}$. The net inventory at the start of the next period is

$$\begin{aligned} I_{n+1} &= I_n - D_n + \min\{c, s - I_n + D_n\} \\ &= \min\{c + I_n - D_n, s\}; \end{aligned} \tag{1}$$

in particular, on-hand inventory never exceeds the target level s.

Our analysis is simplified if, instead of the net inventory I_n, we work with the *shortfall* $Y_n = s - I_n$, the amount by which the target inventory exceeds the net inventory. In light of (1), Y_n is nonnegative and satisfies

$$\begin{aligned} Y_{n+1} &= s - \min\{c + I_n - D_n, s\} \\ &= \max\{Y_n + D_n - c, 0\}. \end{aligned} \tag{2}$$

This is a Lindley recursion and shows that the shortfall sequence coincides with the waiting-time sequence in a single-server queue with service times $\{D_n, n \geq 0\}$ and fixed interarrival time c. This correspondence is used in Tayur and is part of the general treatment of queueing and inventory models in Prabhu (1965).

It follows from (2) that if we assume demands are independent and identically distributed with

$$\mathsf{E}[D_1] < c, \tag{3}$$

then Y_n converges in distribution to a random variable Y satisfying

$$Y \stackrel{d}{=} \max\{Y + D - c, 0\}, \tag{4}$$

where $\stackrel{d}{=}$ denotes equality in distribution and D is a random variable independent of Y having the distribution of demands. Equation (4) thus states that, in stationarity, the shortfalls at the start and end of a period have the same distribution.

Various measures of performance are easily expressed in terms of Y. The long-run average proportion of periods in which no stockout occurs, the *stock availability*, is

$$\alpha(s) = P(Y \leq s). \tag{5}$$

The *fill rate*, which is the long-run average proportion of demands met from stock, is given by

$$\beta(s) = 1 - \frac{\mathsf{E}[\max\{0, \min\{Y + D - c - s, D\}\}]}{\mathsf{E}[D]}. \tag{6}$$

To see this, observe that when the shortfall is Y and the demand is D, the demand not met from stock is all of D if $Y > s$ (no on-hand inventory) and is the amount by which $Y + D - c$ exceeds s if $Y \leq s$. Thus, $\mathsf{E}[\max\{0, \min\{Y + D - c - s, D\}\}]$ is the expected demand not met from stock in each period.

Writing x^+ for $\max\{0, x\}$, the long-run average expected backlog is

$$b(s) = \mathsf{E}[(Y - s)^+]. \tag{7}$$

Dividing this by the mean demand $\mathsf{E}[D]$ gives

$$w(s) = \frac{\mathsf{E}[(Y - s)^+]}{\mathsf{E}[D]}, \tag{8}$$

interpreted as the average delay per unit demand, under the convention that demands filled in the same period they arrive have a delay of 0, demands filled in the next period have a delay of 1, etc. If backorders are penalized at rate $p > 0$ and holding costs charged at rate $h > 0$, then the long-run average cost per period is

$$\begin{aligned} v(s) &= h\mathsf{E}[(s - Y)^+] + p\mathsf{E}[(Y - s)^+] \\ &= h(s - \mathsf{E}[Y]) + (p + h)\mathsf{E}[(Y - s)^+]. \end{aligned} \tag{9}$$

We develop approximations for these performance measures and for base-stock levels that achieve specified values of these measures.

Equation (2) and definitions (5)–(9) presuppose that production decisions are made after the total demand in a period is revealed. To model a system in which production

Table I
The Parameter γ for Some Demand Distributions
(The third column gives defining equations for γ along with the range in which it must lie.)

Name	Density/Mass Fn.	Gamma
Exponential	$\mu e^{-\mu x}$	$\left(\frac{\mu}{\mu-\gamma}\right)e^{-\gamma c}=1$ in $(0,\mu)$
Normal	$(\sigma\sqrt{2\pi})^{-1}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	$\gamma=2(c-\mu)/\sigma^2$
Gamma	$\frac{\mu^m x^{m-1}}{\Gamma(m)}e^{-\mu x}$	$\left(\frac{\mu}{\mu-\gamma}\right)^m e^{-\gamma c}=1$ in $(0,\mu)$
Hyperexpnl.	$p\mu_1 e^{-\mu_1 x}+(1-p)\mu_2 e^{-\mu_2 x}$	$\left[p\left(\frac{\mu_1}{\mu_1-\gamma}\right)+(1-p)\left(\frac{\mu_2}{\mu_2-\gamma}\right)\right]e^{-\gamma c}=1$ in $(0,\min\{\mu_1,\mu_2\})$
Poisson	$e^{-\lambda}\lambda^k/k!$	$\lambda(e^{\gamma}-1)-\gamma c=0,\ \gamma>0$
Neg. Binom.	$\binom{k-1}{m-1}p^m(1-p)^{k-m}$	$m\log[1+p^{-1}(e^{-\gamma}-1)]+\gamma c=0$, in $(0,-\log(1-p))$

must be set before the demand is known, it suffices to introduce a leadtime. See the remark in Section 1.2 and the discussion of leadtimes in Section 3.3 for this case.

1.2. Main Results

Our approximations rely on some mild assumptions on demands. We continue to assume that $\{D_n,\ n\geq 0\}$ are (non-negative and) i.i.d., and we denote their distribution by F_D. Demands are assumed to be either continuous or else integer-valued. In the discrete case, we assume that c and s are also integer-valued and that the demand distribution has unit span. This last assumption is not essential, but it simplifies the discussion. The stability condition (3) is in force throughout. In addition, we assume that

$$P(D_1>c)>0;$$

otherwise, demands can always be met from the current period's production. Our most important assumption involves the moment generating function of D_1-c, given by

$$\phi(\theta)=e^{-\theta c}\int_0^\infty e^{\theta x}\,dF_D(x).$$

We assume that there exists a $\theta_0>0$ at which

$$1<\phi(\theta_0)<\infty. \tag{10}$$

This condition, together with the convexity of ϕ, and the fact that $\phi(0)=1$ and $\phi'(0)=\mathsf{E}[D_1-c]<0$, implies the existence of just one $\gamma>0$ with $\phi(\gamma)=1$, which then has $\phi'(\gamma)<\infty$; that is,

$$\mathsf{E}[e^{\gamma(D_1-c)}]=1, \tag{11}$$

and $\mathsf{E}[(D_1-c)e^{\gamma(D_1-c)}]<\infty$. The solution γ to (11) is called the *conjugate point* for the distribution of D_1-c. It plays a central role in our approximations, beginning with Theorem 1, below. In the statement of the theorem, the notation $f(x)\sim g(x)$ means that $f(x)/g(x)$ converges to unity as $x\to\infty$. For functions of integers, the limit is taken through integer values of x.

Theorem 1. *There is a constant C, $0<C\leq 1$, depending on the demand distribution and the capacity, such that*

(i) *the stock availability satisfies* $1-\alpha(s)\sim Ce^{-\gamma s}$;
(ii) *the average backlog satisfies* $b(s)\sim (C/\gamma)e^{-\gamma s}$;
(iii) *the average delay satisfies* $w(s)\sim (C/\gamma\mathsf{E}[D])e^{-\gamma s}$;
(iv) *the fill rate satisfies* $1-\beta(s)\sim (C/\gamma\mathsf{E}[D])(1-e^{-\gamma c})e^{-\gamma s}$.

The proofs of this theorem and of most of our results are given in Section 4. The conjugate point γ featured in these approximations exists for all commonly used demand distributions; Table I gives examples. Generally, γ is not available explicitly but numerical solution requires little effort. Indeed, in all our examples we evaluated γ to several decimal places of accuracy using basic commands of widely available spreadsheet software and negligible computing time. (See Neuts 1986 for a closely related computation in the matrix-geometric setting; see Whitt for approximations.)

In practice, the constant C may be difficult to evaluate, so further approximation is warranted. (An earlier version of this paper discusses numerical evaluation of C.) A simple upper bound is obtained by replacing C with 1. Somewhat better bounds are obtained by adapting a method of Ross (1974) (see also Asmussen 1993) as follows. Define

$$C_-=\inf_{r\geq c}\left(\mathsf{E}[\exp\{\gamma(D_1-r)\}|D_1>r]\right)^{-1}, \text{ and} \tag{12}$$

$$C_+=\sup_{r\geq c}\left(\mathsf{E}[\exp\{\gamma(D_1-r)\}|D_1>r]\right)^{-1}. \tag{13}$$

Then we have

Theorem 2. *For all* $s > 0$,

(i) $C_- e^{-\gamma s} \leq 1 - \alpha(s) \leq C_+ e^{-\gamma s}$;
(ii) $(C_-/\gamma)e^{-\gamma s} \leq b(s) \leq (C_+/\gamma)e^{-\gamma s}$;
(iii) $(C_-/\gamma\mathsf{E}[D])e^{-\gamma s} \leq w(s) \leq (C_+/\gamma\mathsf{E}[D])e^{-\gamma s}$;
(iv) $(C_-/\gamma\mathsf{E}[D])[1 - e^{-\gamma c}]e^{-\gamma s} \leq 1 - \beta(s) \leq (C_+/\gamma\mathsf{E}[D])[1 - e^{-\gamma c}]e^{-\gamma s}$.

We give some explicit expressions for C_- and C_+ and illustrate their use in Section 2.

As a special case, consider exponentially distributed demands with mean $1/\mu$. In this setting, as noted by Tayur, the stationary shortfall distribution coincides with the stationary waiting time in a D/M/1 queue. Moreover, by the memoryless property, the conditional expectations in (12) and (13) do not depend on r, so $C_- = C_+ = 1 - \gamma/\mu$ and

$$P(Y > s) = \left(1 - \frac{\gamma}{\mu}\right)e^{-\gamma s}, \tag{14}$$

where γ solves $\mu e^{-\gamma c}/(\mu - \gamma) = 1$. Indeed, the equality $P(Y > s) = Ce^{-\gamma s}$ holds for all G/M/1 queues; see Prabhu (1965, p. 109) or Asmussen (1987, p. 204).

As an application of Theorems 1 and 2, we consider the problem of setting the base-stock level to ensure that, over an infinite horizon, stockouts occur in at most a fraction $\delta > 0$ of periods. This is the problem of setting s so that $\alpha(s) \geq 1 - \delta$, with δ equal to, say, 0.01.

Corollary 1. *Let* s_δ *be the minimal base-stock level for which a stock availability of at least* $1 - \delta$ *is guaranteed; i.e., the minimal* s *satisfying* $\alpha(s) \geq 1 - \delta$. *Then*

$$\gamma^{-1}\log(C_-/\delta) \leq s_\delta \leq \gamma^{-1}\log(C_+/\delta) \text{ for all sufficiently small } \delta > 0, \tag{15}$$

and $s_\delta \leq -\gamma^{-1}\log\delta$ *for all* $\delta > 0$. *If the demand distribution is continuous,*

$$|s_\delta - \gamma^{-1}\log(C/\delta)| \to 0 \quad \text{as } \delta \to 0. \tag{16}$$

Let s'_δ *be the smallest* s *for which* $\beta(s) \geq 1 - \delta$. *Then*

$$\gamma^{-1}\log(C_-[1 - e^{-\gamma c}]/\gamma\delta\mathsf{E}[D_1]) \leq s'_\delta \leq \gamma^{-1}\log(C_+[1 - e^{-\gamma c}]/\gamma\delta\mathsf{E}[D_1]), \tag{17}$$

for all sufficiently small $\delta > 0$. *If demands are continuous, then*

$$|s'_\delta - \gamma^{-1}\log(C[1 - e^{-\gamma c}]/\gamma\delta\mathsf{E}[D_1])| \to 0 \quad \text{as } \delta \to 0. \tag{18}$$

Proof. The bounds on s_δ follow from Theorem 2(i) by inverting the bounds on α and the fact that $C \leq 1$. The bounds in Theorem 2 are valid only for $s > 0$, so to invert them we need $\delta < C_+$ in (15) and $\delta < C_+[1 - e^{-\gamma c}]/(\gamma\mathsf{E}[D_1])$ in (17), which is not a restriction since we are primarily interested in small δ.

For (16), notice from (15) that $s_\delta \to \infty$ as $\delta \to 0$, so from Theorem 1(i), $C^{-1}e^{\gamma s_\delta}P(Y > s_\delta) \to 1$; i.e., $C^{-1}e^{\gamma s_\delta}\delta \to 1$, implying that

$$|\log C^{-1} + \gamma s_\delta + \log\delta| \to 0,$$

which is equivalent to (16). The assertions regarding s'_δ are proved in the same way. □

With a discrete demand distribution, we may have $P(Y > s_\delta) < \delta$ for arbitrary δ, but we still have $P(Y > s_{\delta_n}) = \delta_n$ through a subsequence $\{\delta_n\}$, so

$$\liminf_{\delta \to 0} |s_\delta - \gamma^{-1}\log(C/\delta)| = 0; \tag{19}$$

moreover, the difference never exceeds 1.

A variant of Corollary 1 holds for the base-stock level minimizing the long-run average cost $v(s)$ defined in (9). We examine the optimal base-stock level as the backorder penalty p becomes large.

Corollary 2. *Suppose* F_D *is continuous; then* v *is convex. Suppose* s_p *minimizes* v; *then*

$$\gamma^{-1}\log((p + h)C_-/h) \leq s_p \leq \gamma^{-1}\log((p + h)C_+/h) \text{ for all sufficiently large } p, \tag{20}$$

and $s_p \leq \gamma^{-1}\log((p + h)/h)$ *for all* $p > 0$. *Moreover,*

$$|s_p - \gamma^{-1}\log((p + h)C/h)| \to 0 \quad \text{as } p \to \infty. \tag{21}$$

Proof. Differentiation of (9) yields $v'(s) = h - (p + h)P(Y > s)$, an increasing function of s, making v convex. The minimum of v is achieved at the point s_p satisfying $v'(s_p) = 0$; i.e., satisfying $P(Y > s_p) = h/(p + h)$. The bounds and limiting behavior of s_p thus follow from the bounds and asymptotics of the tail distribution of Y, just as in Corollary 1. □

Remarks. (i) The conclusions in (16), (18), and (21) are very strong. It is clear, for example, that s_p must increase as the penalty p increases. A result of the form $s_p \sim \gamma^{-1}\log((p + h)C/h)$ would imply that s_p increases at the same rate as the approximating logarithm. But (21) shows not only that the two expressions increase at the same rate, but that the difference between them vanishes as p increases.

(ii) The limits in (16) and (21) are taken as $\delta \to 0$ and $p \to \infty$, respectively. One cannot also let $c \to \infty$ without further justification; thus, our asymptotics do not necessarily recover results for uncapacitated systems as the capacity becomes large. (This should be kept in mind in later sections as well.) In general, the effectiveness of our bounds and approximations tends to increase with the ratio $\rho = \mathsf{E}[D]/c$.

As already noted, in the case of exponentially distributed demands the approximation $P(Y > s) \sim C\exp(-\gamma s)$ becomes exact. Accordingly, the optimal s_p becomes exactly $\gamma^{-1}\log((p + h)C/h)$ for all $p > 0$. This is Tayur's result, except for the fact that he penalizes backorders after demands arrive, before the current period's production. In other words, he solves $P(Y + D > s) = h/(p + h)$, rather than $P(Y > s) = h/(p + h)$. This can also be viewed as introducing a leadtime of 1 between production and availability of finished goods. Our approximations are easily adapted to this case; in particular, for $s \geq c$,

$$P(Y + D > s) = P(\max\{Y + D - c, 0\} > s - c)$$
$$= P(Y > s - c) \sim Ce^{-\gamma(s-c)},$$

so the effect on the asymptotic behavior is to change the constant C to $Ce^{\gamma c}$. It follows that in (15) and (20), c should be added to both the lower and upper bounds (with the lower bound now valid for $s > c$, rather than $s > 0$), and in (16) and (21), c should be added to the limiting approximations.

We give further exact expressions for the exponential case in the following result. These are direct consequences of (14).

Proposition 1. *Suppose the demand distribution is* $F_D(x) = 1 - \exp[-\mu x]$. *Then, for all* $s > 0$,

(i) $1 - \alpha(s) = (1 - \gamma/\mu)e^{-\gamma s}$;
(ii) $b(s) = (1 - \gamma/\mu)\gamma^{-1}e^{-\gamma s}$;
(iii) $1 - \beta(s) = e^{-\gamma(s+c)}$.

1.3. Imperfect Production

The approximations of the previous section can be modified to account for variability in production resulting from yield losses, variability in capacity, or shortages of raw material. A simple model of imperfect production replaces the fixed capacity c with an i.i.d. sequence $\{Z_n, n \geq 1\}$ in which Z_n represents the maximum nondefective production in period n. With this modification, the shortfall evolution becomes

$$Y_{n+1} = \max\{0, Y_n + D_n - Z_n\},$$

keeping us within the general framework of the previous section. Suppose $\mathsf{E}[D_1] < \mathsf{E}[Z_1]$ and $P(D_1 > Z_1) > 0$, and suppose the moment generating function of $D_1 - Z_1$ satisfies (10). Then there is a $\gamma > 0$ solving

$$\mathsf{E}[\exp\{\gamma(D_1 - Z_1)\}] = 1, \tag{22}$$

and the argument that proves Theorem 1 shows that there is a constant C (depending on the distributions of D_1 and Z_1) such that

$$P(Y > x) \sim Ce^{-\gamma x},$$

which is equivalent to part (i) of Theorem 1. The other results of Section 1.2 follow accordingly.

This extension can be used to assess the impact of production variability on required base-stock levels. In the simplest case, the production facility is down with probability q and up with probability $1 - q$. To keep the average capacity fixed at c, we thus set

$$P(Z_1 = 0) = q, \quad P(Z_1 = c/(1 - q)) = 1 - q. \tag{23}$$

With γ_q the corresponding solution to (22), we examine the dependence of γ_q on q since this, then, determines the dependence of base-stock levels on the failure probability. An exact analysis is possible in the case of exponentially distributed demands:

Lemma 1. *Suppose the demand distribution is* $F_D(x) = 1 - \exp[-\mu x]$. *Then* $\gamma_q = (1 - q)\gamma$, *where* $\gamma = \gamma_0$ *is the parameter in the case of perfect production.*

Proof. Under (23) and exponential demands, Equation (22) becomes

$$1 = \mathsf{E}[\exp(\gamma_q D_1)](q + (1 - q)\exp[-\gamma_q c/(1 - q)])$$
$$= (\mu/(\mu - \gamma_q))(q + (1 - q)\exp[-\gamma_q c/(1 - q)]). \tag{24}$$

At $q = 0$, this simplifies to

$$1 - e^{-\gamma c} - \frac{\gamma}{\mu} = 0. \tag{25}$$

Since (24) has at most one nonzero solution in $(0, \mu)$, it suffices to show that setting $\gamma_q = (1 - q)\gamma$ solves (24). Making this substitution and collecting terms involving q yields

$$q\left(1 - e^{-\gamma c} - \frac{\gamma}{\mu}\right) = 1 - e^{-\gamma c} - \frac{\gamma}{\mu},$$

which holds, in light of (25). □

Thus, in the case of exponential demands, the bounds $s_\delta \leq -\gamma^{-1}\log\delta$ and $s_p \leq -\gamma^{-1}\log[h/(p + h)]$ (which also serve as rough approximations) become

$$s_\delta \leq -((1 - q)\gamma)^{-1}\log\delta, \quad \text{and}$$
$$s_p \leq -((1 - q)\gamma)^{-1}\log[h/(p + h)],$$

to account for the failure probability q. Taking this a step further, we obtain exact results:

Proposition 2. *Suppose the demand distribution is* $F_D(x) = 1 - \exp[-\mu x]$. *Then*

$$s_\delta = -((1 - q)\gamma)^{-1}\log(\delta/C_q), \quad \text{and}$$
$$s_p = -((1 - q)\gamma)^{-1}\log\left(\frac{h}{C_q(p + h)}\right), \quad \text{where}$$
$$C_q = 1 - \frac{\gamma_q}{\mu} = 1 - \frac{(1 - q)\gamma}{\mu},$$

and $\gamma = \gamma_0$ *is the conjugate point for the case of perfect production.*

For more general demand distributions, it does not seem possible to give an explicit expression for the dependence of γ_q on q. But it is a simple matter to examine this dependence numerically. Figure 1 illustrates this dependence for various Erlang and hyperexponential distributions with fixed mean. (The graphs are piecewise linear interpolations of points calculated at the q-values labeled on the horizontal axis.) Decreasing the coefficient of variation below 1 appears to make γ_q more convex in q; increasing it above 1 appears to make the dependence more concave. In addition, the impact of demand variability is most pronounced at low failure probabilities.

Explicit results on the effect of production variability are also possible if we postulate normal distributions for both demand and production. Any normal distribution assigns

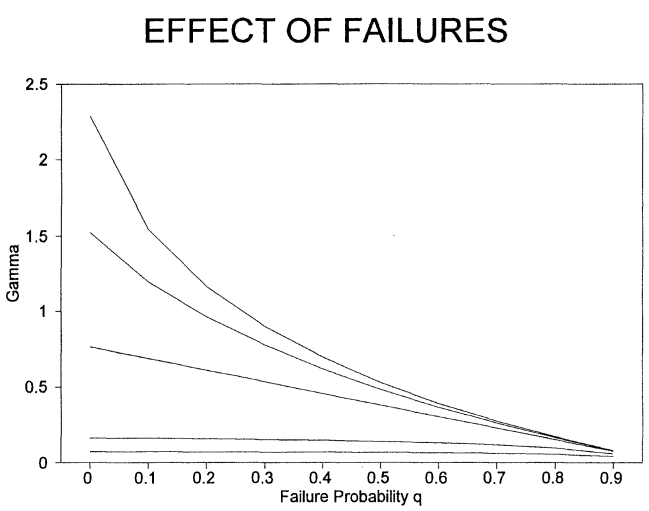

Figure 1. The dependence of γ_q on the failure probability q for five demand distributions. In each case, $c = 1$ and the mean demand is 0.7. From top to bottom, the curves correspond to the following demand distributions: 3-stage Erlang, 2-stage Erlang, exponential, and hyperexponentials with cv = 2 and cv = 3.

positive probability to negative values, but this probability is made negligible by a sufficiently small variance. Let us suppose, then, that demands are normal with mean μ and variance σ^2, and that maximum production is normal with mean c and variance σ_c^2. Denote by γ_{σ_c} the conjugate point when the variance of capacity is σ_c^2. A simple manipulation of moment generating functions (see Table I) proves the following:

Proposition 3. *For normally distributed demands and capacity, the effect of variability in capacity is to set* $\gamma_{\sigma_c} = (\sigma^2/(\sigma^2 + \sigma_c^2))\gamma$ *where* σ^2 *is the demand variance and* $\gamma \equiv \gamma_0$ *is the conjugate point for constant capacity.*

Thus, in the normal case, γ is a decreasing, convex function of production variability—the same qualitative dependence as observed on q in Figure 1 for Erlang demand (but not for hyperexponential demand which appears concave). The normal case lends itself to accurate numerical approximation of the constant C, making it attractive for calculation; specifically, adapting page 175 of Siegmund, one obtains the approximation

$$C \approx \exp\left[-2(0.583)\left(\frac{c-\mu}{\sqrt{\sigma^2+\sigma_c^2}}\right)\right].$$

Combining this with Proposition 3 and (21), we arrive at the approximation

$$s_p \approx \left(\frac{\sigma^2+\sigma_c^2}{2(c-\mu)}\right)\log\left(\frac{p+h}{h}\right) - 0.583\sqrt{\sigma^2+\sigma_c^2}, \quad (26)$$

suggesting that the optimal base-stock level is convex in σ_c^2.

2. COMPUTING BOUNDS

In this section, we evaluate the bounds in Theorem 2 for a variety of demand distributions, including the Erlang and hyperexponential families. These families provide two-moment approximations to all distributions and can thus model a wide range of demand patterns. We show through examples that the bounds for the Erlang and hyperexponential families are effective.

In referring to demand distributions, we use some abbreviations. We say that demand is $E_m(\mu)$ if

$$F_D(x) = 1 - \sum_{i=0}^{m-1} \frac{(\mu x)^i}{i!} e^{-\mu x},$$

and demand is $H_2(\mu_1, \mu_2, p)$ if

$$F_D(x) = 1 - pe^{-\mu_1 x} - (1-p)e^{-\mu_2 x},$$

with $\mu_1 \leq \mu_2$ and $0 \leq p \leq 1$. Demands follow an NBU distribution (new better than used) if

$$1 - F_D(x+y) \leq (1 - F_D(x))(1 - F_D(y)), \quad \text{for all } x, y \geq 0,$$

and NWU (new worse than used) if the reverse inequality holds; see Chapter 6 of Barlow and Proschan (1975) for background. This terminology arises in reliability theory; the terms *new* and *used* have no interpretation for demand random variables, but the classes of distributions NBU and NWU are still useful in modeling.

We summarize bounds on C in

Proposition 4. *The following bounds hold:*

(i) *For all demand distributions,* $C \leq 1$.

(ii) *If demands are NBU then* $C_- = e^{-\gamma c} \leq C$, *and if demands are NWU then* $C \leq e^{-\gamma c} = C_+$.

(iii) *If demands are* $E_m(\mu)$, *then*

$$C_- = e^{-\gamma c} \leq C \leq e^{-\gamma c/m} = C_+.$$

The same is true for the negative binomial distribution in Table I.

(iv) *If demands are* $H_2(\mu_1, \mu_2, p)$, *then*

$$C_- = 1 - \frac{\gamma}{\mu_1} \leq C \leq e^{-\gamma c} = C_+.$$

(v) *If demands are bounded above by b, then* $e^{-\gamma b} \leq C$.

Figures 2–4 illustrate the performance of these bounds in a range of settings. Figure 2 plots the upper and lower bounds on s_δ given in (15) as the stock availability $1 - \delta$ ranges from 90% to nearly 100%. The bounds plotted are for two-stage Erlang demand with a mean of 0.9. In all our examples we take $c = 1$, meaning that s is measured in units of capacity. Thus, the ratio $\mathsf{E}[D_1]/c$, which we call ρ, is just the mean demand. The graph shows that the bounds are quite close together. Indeed, it follows from (15) and Proposition 4 that the vertical distance between the two curves is $\gamma^{-1}\log(C_+/C_-) = 1/2$ for all δ. With m-stage

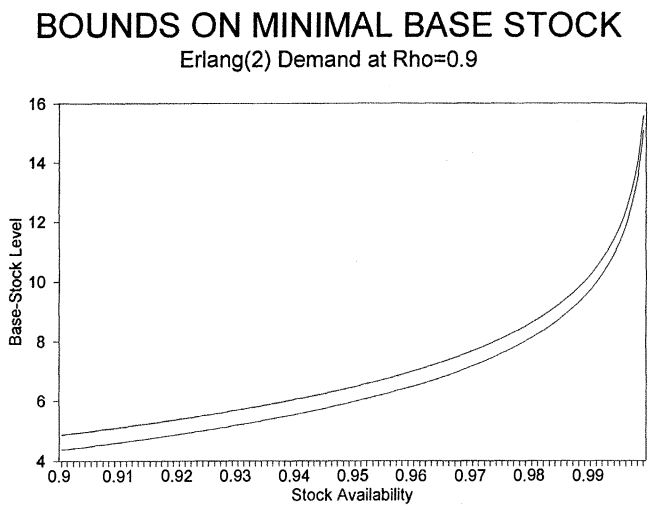

Figure 2. Upper and lower bounds on the minimal base-stock level required to meet a specified stock availability. Demand in each period has a 2-stage Erlang distribution. The ratio ρ of mean demand to capacity is 0.9.

Erlang demands, the vertical gap is $1 - 1/m$, so the gap increases with m.

Based on (20), Figure 3 presents similar results for the optimal base-stock level as a function of the ratio p/h of backorder penalty to holding cost. The lower two curves give bounds for $\rho = 0.7$, the upper two curves are for $\rho = 0.9$. The graph illustrates that the bounds are quite close compared to the change in s_p when ρ is increased from 0.7 to 0.9. Not surprisingly, the optimal base-stock level increases with ρ. As in Figure 2, the vertical distance between each pair of bounds is 1/2, and would be $1 - 1/m$

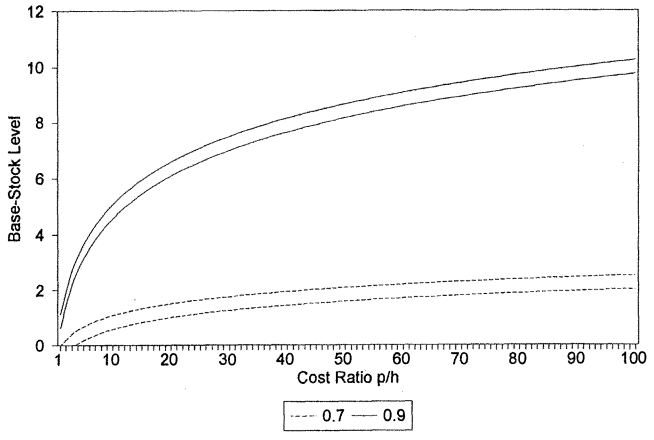

Figure 3. Upper and lower bounds on optimal base-stock levels for various ratios of backorder penalty to holding cost. Demand in each period has a 2-stage Erlang distribution. The ratio ρ of mean demand to capacity is 0.9 in the higher pair of curves and 0.7 in the lower pair.

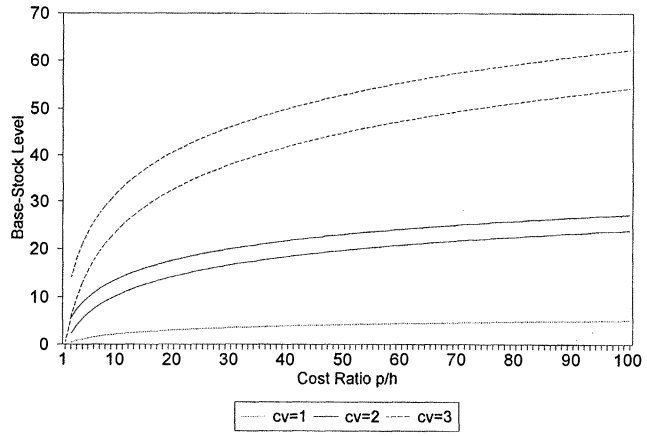

Figure 4. Upper and lower bounds on optimal base-stock levels at $\rho = 0.7$ and three different coefficients of variation (cv). Demand in each period has a hyperexponential distribution. The highest pair of curves are bounds for cv = 3; the middle pair are bounds for cv = 2; and the lowest curve is the exact value for cv = 1 (exponential demands).

with m-stage Erlang demands. It can be shown using Theorem 2 and (20) that the additional cost incurred by setting s at either the upper or lower bound (instead of the optimum) is bounded in p, and thus the *relative* loss of optimality vanishes as p increases. For the cases in Figure 3, we found through simulation that the cost gap using the lower bound ranges from about 10% to less than 1%, whereas the cost at the upper bound is virtually indistinguishable from the optimum.

Figure 4 shows results for hyperexponential demand. We fix ρ at 0.7 and examine the impact of demand variability, as measured by the coefficient of variation (cv). A cv of 1 corresponds to an exponential distribution for which we have exact results; hence, there is only one curve for that case. The other cases (cv = 2, cv = 3) show the dramatic increase in base stock necessitated by increased variability. For each cv, the vertical distance between the upper and lower bounds is constant; that vertical distance increases with the cv because γ decreases to zero as the cv increases. The impact on cost is again bounded independent of p. For the cases in Figure 4, we found through simulation that the cost gap using the upper bound ranges from about 3% to 8%, whereas the cost at the lower bound is virtually indistinguishable from the optimum.

It is possible, in principle, to supplement Figures 2 and 3 with numerical approximations to the exact values of s_δ and s_p using a result in Prabhu (1965, p. 217) for $D/E_m/1$ queues. More generally, Ramaswami and Lucantoni (1985) give a procedure for computing tail probabilities of the stationary waiting time in G/PH/1 queues and this

could be used for phase-type demand in our setting. However, these methods require substantially more work than our bounds and approximations.

3. MULTISTAGE SYSTEMS

We now consider multistage systems consisting of d nodes in series. Node 1 supplies external demands, node i draws material from node $i+1$, $i = 1, \ldots, d-1$, and node d draws from an unlimited supply of raw material. Node i has capacity c^i, $i = 1, \ldots, d$.

3.1. Shortfall Formulation

Each node of the multi-stage system follows a base-stock policy for *echelon* inventory, operating as follows. Let I_n^1 be the net inventory (stock on hand minus backorders) at stage 1, and let I_n^i be the inventory available at stage i, $i = 2, \ldots, d$, all in period n. The echelon-i inventory at the start of period n is $I_n^1 + \ldots + I_n^i$, $i = 1, \ldots, d$; this drops by D_n upon the arrival of demands in that period. Subsequently, stage i sets production to restore the echelon-i inventory to a base-stock level s^i, while not exceeding its capacity c^i or the available supply I_n^{i+1} of predecessor inventory. Thus, production at node i in period n is given by

$$\min\left\{s^i + D_n - \sum_{j=1}^{i} I_n^j,\ c^i,\ I_n^{i+1}\right\}, \tag{27}$$

with $I_n^{d+1} \equiv \infty$ since node d is not constrained by upstream supply. Because the s^i's are target levels for *cumulative* stock, we always assume that $s^1 \leq s^2 \leq \cdots \leq s^d$.

As shown in Glasserman and Tayur (1994), the dynamics of this system are conveniently represented through *echelon shortfalls*. The shortfall for echelon i at the start of period n is

$$Y_n^i = s^i - \sum_{j=1}^{i} I_n^j,$$

the difference between the target and actual echelon inventories. Using the expression in (27) for the production at stage i, Glasserman and Tayur (1994) show that the shortfalls satisfy

$$Y_{n+1}^d = \max\{0,\ Y_n^d + D_n - c^d\}; \tag{28}$$

$$Y_{n+1}^i = \max\{0,\ Y_n^i + D_n - c^i,\ Y_n^{i+1} + D_n - (s^{i+1} - s^i)\}, \quad i = 1, \ldots, d-1. \tag{29}$$

We show how these recursions lead to approximations.

3.2. Main Results

We continue to assume that demands are i.i.d. with distribution F_D. With $Y_n = (Y_n^1, \ldots, Y_n^d)$, Glasserman and Tayur (1994) show that the process $\{Y_n, n \geq 1\}$ admits a finite stationary distribution to which it converges from all initial distributions, provided

$$\mathsf{E}[D_1] < c^* \equiv \min\{c^1, \ldots, c^d\}. \tag{30}$$

For our approximations, we require that $P(D_1 > c^*) > 0$ and that there exists a $\theta_0 > 0$ at which

$$1 < \mathsf{E}[e^{\theta_0(D_1 - c^*)}] < \infty.$$

We denote by γ the unique nonzero solution to

$$\mathsf{E}[e^{\gamma(D_1 - c^*)}] = 1.$$

Our results are easiest to formulate under the additional assumption that there is just one stage i^* with capacity c^* and thus that all other stages have strictly greater capacity. Later, we remove this condition. With i^* as just defined, let

$$\eta = \min_{j \geq i^*}\{(s^j - s^1) - (j-1)c^*\}. \tag{31}$$

This quantity provides the link between the analysis of single-stage and multistage systems. In the important special case that $i^* = d$ (meaning that the bottleneck stage is highest in the hierarchy), we have simply

$$\eta = (s^d - s^1) - (d-1)c^*. \tag{32}$$

We now have the following multistage version of Theorem 1(i):

Theorem 3. *Let $Y = (Y^1, \ldots, Y^d)$ have the stationary distribution of the shortfall process. Then*

$$P(Y^1 > x) \sim Ce^{-\gamma(x+\eta)} \quad \text{as } x \to \infty, \tag{33}$$

where C is the constant for a single-stage system with capacity c^* and the same demand distribution as the multistage system.

Thus, the tail distribution of stock to serve external demands in a multistage system corresponds to that in a single-stage system with the minimal capacity, except that the constant C is replaced by $C\exp(-\gamma\eta)$. In particular, when $i^* = d$, tail probabilities ultimately depend only on c^d and $s^d - s^1$. Since Theorem 3 then suggests that the probability of a stockout admits the approximation $P(Y^1 > s^1) \approx C\exp[-\gamma(s^d - (d-1)c^*)]$, it further suggests that the dominant features determining the ability to meet demands are the minimal capacity and the system-wide base-stock level s^d.

For any $s = (s^1, \ldots, s^d)$, the stock availability, the fill rate and the average backorders are defined for the multistage system just as for the single stage system but replacing Y with Y^1 in (5), (6) and (7). From Theorem 3 we get

Corollary 3. *Parts* (i)–(iv) *of Theorem* 1 *hold for multistage systems, with C replaced by $C\exp(-\gamma\eta)$ and c replaced by c^*.*

These approximations can also be formulated with respect to a single-stage system. For example, if $\alpha^*(s)$ is the stock availability for a single-stage system with capacity c^*, then

$$1 - \alpha(s) \sim e^{-\gamma\eta}(1 - \alpha^*(s)).$$

The interpretation of these results is in some cases clearer if we express the approximations in terms of the *incremental* base-stock levels $\Delta^i = s^i - s^{i-1}$, $i = 2, \ldots, d$. With this notation, we have

$$\eta = \min_{j \geq i^*} \left\{ \left(\sum_{i=2}^{j} \Delta^i \right) - (j-1)c^* \right\},$$

an expression not depending explicitly on s^1. Now fix $\Delta^2, \ldots, \Delta^d$ and let s_δ^1 be the corresponding minimal stage-1 base-stock level required to guarantee a stock availability of at least $1 - \delta$ for any $0 < \delta < 1$. Then we have

Corollary 4. *For any $\Delta^2, \ldots, \Delta^d$ and any $0 < \delta < 1$, $s_\delta^1 \leq -\gamma^{-1} \log \delta - \eta$. If the demand distribution is continuous, then*

$$|s_\delta^1 - [\gamma^{-1} \log(C/\delta) - \eta]| \to 0, \quad as\ \delta \to 0.$$

Corollary 4 can be supplemented with bounds, much as in Section 1.2, using $C_- e^{-\gamma\eta_+}$ and $C_+ e^{-\gamma\eta_-}$ in place of C_-, C_+, with η_-, η_+ defined in (48). Approximating the base-stock level s_p^1 that minimizes system cost is more difficult, because holding costs are typically charged on inventory at all stages, not just the lowest stage. Hence, the optimal s_p^1 is no longer characterized by a simple condition on the tail distribution of Y^1. One might, however, choose to set base-stock levels by ensuring a certain stock availability at upper echelons and minimizing holding and penalty costs at stage 1 subject to that constraint on $s^2, \ldots, s^d$. For that approach, tail probabilities for $Y^2, \ldots, Y^d$ are relevant; we approximate these next.

For $k = 1, \ldots, d$, let i_k^* be the index of the stage with the smallest capacity among those in $\{k, k+1, \ldots, d\}$, which we assume is unique. (This holds if, e.g., no two capacities are equal.) Let c_k^* be the capacity of stage i_k^*. Define

$$\eta_k = \min_{j \geq i_k^*} \left\{ \sum_{i=k+1}^{j} \Delta_i - (j-k)c_k^* \right\},$$

and notice that η_k coincides with the η in (31) for the subsystem consisting of stages $k, k+1, \ldots, d$. Since the evolution of Y_n^k is unaffected by that of Y_n^i, $i = 1, \ldots, k-1$, a consequence of Theorem 3 is this:

Theorem 4. *For all $k = 1, \ldots, d$,*

$$P(Y^k > x) \sim C_k \exp[-\gamma_k(x + \eta_k)],$$

where γ_k solves $\mathsf{E}[\exp\{\gamma_k(D_1 - c_k^*)\}] = 1$ and C_k is the constant C for a single-stage system with capacity c_k^*.

We can remove the requirement that there be just one stage with the minimal capacity c^* (or c_k^*) by making a simple modification. If, say, stages $i_1, \ldots, i_m$ all have capacity c^*, then η is determined by the lowest stage; that is, we have

$$\eta = \min_{j \geq i_1} \{(s^j - s^1) - (j-1)c^*\}. \tag{34}$$

The definition of η_k is modified analogously.

Because the distribution of each Y^k potentially depends on all of $\Delta^{k+1}, \ldots, \Delta^d$, Theorem 4 does not provide a direct solution to the problem of setting base-stock levels. However, it does provide a basis for approximating the distribution of shortfalls, and thus also for approximating costs. The simplest approximation sets

$$P(Y^1 > x) \approx Ce^{-\gamma(x+\eta)}, \quad \mathsf{E}[Y^1] \approx Ce^{-\gamma\eta}/\gamma,$$
$$\mathsf{E}[(Y^1 - s^1)^+] \approx Ce^{-\gamma(s^1+\eta)}/\gamma, \tag{35}$$

and similarly for $Y^2, \ldots, Y^d$. If echelon-i inventory is charged a holding cost at rate h_i and if backorders are penalized at rate p, the long-run average cost per period becomes

$$\sum_{i=1}^{d} h_i(s^i - \mathsf{E}[Y^i]) + (p + h_1 + \cdots + h_d)$$
$$\mathsf{E}[(Y^1 - s^1)^+],$$

which generalizes (9). Substituting for the expectations according to (35) results in a cost approximation. Corresponding upper and lower bounds follow from replacing C with $C_+ e^{-\gamma\eta_-}$ and $C_- e^{-\gamma\eta_+}$, with C_-, C_+ as in Section 1.2 and η_-, η_+ as in (48).

The performance of these bounds and approximations is illustrated for a two-stage system in Table II; the approximation above is labeled "Approx1." A shortcoming of this simple approximation is that it is insensitive to all capacities except the smallest. A modification developed in Glasserman and Tayur (1996) uses

$$P(Y^1 > x) \approx (1 - \exp(-\gamma[(s^2 - s^1) - c_1]^+))C'e - \gamma' x + Ce^{-\gamma(x+\eta)},$$

where C', γ' are the constants for stage 1 viewed as a single-stage system in isolation. This approximation is consistent with Theorem 3. Its performance is illustrated in the table under "Approx2"; it generally has smaller error than the straightforward approximation. A numerical study of this approximation, generalized to five-node systems, is presented in Glasserman and Tayur (1996). As with our earlier results, the quality of these approximations appears to increase with ρ.

3.3. Systems with Leadtimes

Thus far, we have assumed that period-n production at stage $i+1$ becomes available input to stage i in period $n+1$, for all n and all $i = 1, \ldots, d-1$. We now consider a modification in which there are fixed, exogenous leadtimes for each stage; these could model transportation times between stages, for example. Cumulative leadtimes are specified by positive integers $l^1 < l^2 < \cdots < l^d$ as follows. Stage-1 production becomes available to meet external demands after l^1 periods; stage-$(i+1)$ production becomes available input to stage-i after $l^{i+1} - l^i$ periods, $i = 1, \ldots, d-1$. Thus, l^i is the total leadtime from stage i to external demands. Our previous model had $l^i = i$, $i = 1, \ldots, d$. The natural counterpart to (34) is

$$\eta = \min_{j \geq i_1} \{(s^j - s^1) - (l^j - 1)c^*\}. \tag{36}$$

With this definition, we have

Table II
Performance of Bounds and Approximations in a Two-Stage System

	$s^2 - s^1$	Simulation	Lower	Upper	Approx1	Approx2
$c^1 = 1$	1	8.17 (0.169)	8.16	8.16	8.16	8.16
	1.3	8.47 (0.169)	7.54	8.71	7.79	8.46
	1.8	8.97 (0.169)	6.91	9.52	7.47	8.96
	2.5	9.67 (0.169)	6.60	10.5	7.43	9.66
$c^1 = 1.5$	1	8.17 (0.169)	8.16	8.16	8.16	8.16
	1.3	7.80 (0.147)	7.54	8.71	7.79	7.79
	1.8	7.49 (0.115)	6.91	9.52	7.47	7.52
	2.5	7.49 (0.080)	6.60	10.5	7.43	7.57
$c^1 = 2$	1	8.17 (0.361)	8.16	8.16	8.16	8.16
	1.3	7.80 (0.147)	7.54	8.71	7.79	7.79
	1.8	7.48 (0.114)	6.91	9.52	7.47	7.47
	2.5	7.44 (0.080)	6.60	10.5	7.43	7.45

Demands are exponential with mean 0.7; $c^2 = 1$ and $s^1 = 1.5$. Cost parameters are $h_1 = 2$, $h_2 = 1$, and $p = 20$. Numbers in parentheses are 95%-confidence-interval halfwidths.

Theorem 5. *In a multistage system with cumulative leadtimes $l^1, \ldots, l^d$, the stage-1 shortfall satisfies* (33) *with γ the solution to* $\mathsf{E}[\exp\{\gamma(D_1 - c^*)\}] = 1$, *$C$ the constant for a single-stage system with capacity c^*, and η as in* (36). *In particular, if the capacity at stage d is strictly less than that at any other stage, then*

$$P(Y^1 > x) \sim C \exp[-\gamma\{(s^d - s^1) - (l^d - 1)c^d\}]e^{-\gamma x}.$$

Proof. It is shown in Glasserman and Tayur (1994) that the evolution of Y_n^1 in a system with fixed leadtimes is identical to that of Y_n^1 in a system with $l^{i+1} - l^i - 1$ dummy nodes between production facilities $i + 1$ and i and unit leadtimes throughout. The dummy nodes between facilities $i + 1$ and i all have capacity c^{i+1} and (echelon) base-stock level s^{i+1}. Their effect is to advance stage-$(i + 1)$ production by one node each period, thereby mimicking the effect of the leadtimes. Since the system with dummy nodes has unit leadtimes, Theorem 3 applies to it. For this modified system, the η in (34) is given by the η in (36). □

We can apply the reduction of a system with leadtimes to one with dummy nodes to the case of a single-stage system with leadtime $l > 1$. The shortfall process Y in such a model coincides with the shortfall process Y^1 in an l-stage system with $c^i \equiv c$ and $s^i \equiv s$, $i = 1 \ldots, l$. It follows that $P(Y > x) \sim C \exp[-\gamma(x - (l - 1)c)]$, and in particular that the stock availability satisfies

$$1 - \alpha(s) \sim Ce^{\gamma(l-1)c}e^{-\gamma s}.$$

Thus, the base-stock level s_δ required for a stockout probability not exceeding δ satisfies

$$|s_\delta - [\gamma^{-1}\log(\delta/C) + (l - 1)c]| \to 0. \tag{37}$$

A comparison with (16) reveals that, asymptotically, the effect of the leadtime l is to increase the required base-stock level by $(l - 1)c$.

This result can be understood intuitively as follows. When s is large, stockouts occur only following several periods of large demand. In each period in which demand is large (at least as large as the capacity), the amount produced (and thus added to pipeline inventory) is c. Following several periods of large demand, the total inventory in transit is therefore $(l - 1)c$. So, asymptotically, $(l - 1)c$ is the amount by which the inventory immediately available to meet demands is less than the total inventory on hand or in transit. Increasing s by $(l - 1)c$ compensates for this deficit to maintain the stockout probability at δ. Equation (37) holds for fixed c with $\delta \to 0$; like our other results, it cannot be applied for fixed δ with $c \to \infty$.

4. THEORETICAL DEVELOPMENTS

In this section, we first review some necessary background on random walks, then give proofs of our main results.

4.1. A Random Walk and Its Conjugate

Let $X_n = D_n - c$ and $S_n = X_1 + \cdots + X_n$, for all $n \geq 1$ with $S_0 = 0$. Reflecting this random walk at the origin yields the shortfall process; moreover, if we let $M_n = \max_{1 \leq i \leq n} S_i$, then a classical result states that Y_n and M_n have the same distribution (see, e.g., Prabhu 1980, §1.5). Since $\mathsf{E}[X_1] = \mathsf{E}[D_1] - c < 0$, the maximum over all time $M = \max_n S_n$ is finite with probability one and has the same distribution as Y. Thus, tail probabilities for Y can be analyzed as tail probabilities for M.

From the demand distribution F_D define a new distribution $\tilde{F}$ on $(-c, \infty)$ by

$$\tilde{F}(x) = \int_0^{x+c} e^{\gamma(t-c)}\, dF_D(t), \quad x > -c. \tag{38}$$

Condition (11) ensures that $\tilde{F}$ is indeed a probability distribution. Let $\{\tilde{X}_n, n \geq 0\}$ be i.i.d. with distribution $\tilde{F}$ and set

$$\tilde{S}_n = \tilde{X}_1 + \cdots + \tilde{X}_n, \quad n \geq 1; \quad \tilde{S}_0 = 0. \tag{39}$$

This is the *conjugate* random walk associated with $\{S_n, n \geq 0\}$. A simple calculation shows that

$$\mathsf{E}[\tilde{X}_1] = \phi'(\gamma) > 0, \tag{40}$$

so $\tilde{S}_n$ has positive drift, whereas S_n has drift $\mathsf{E}[D_1 - c] < 0$.

We now analyze $P(Y > x)$, paralleling the treatment in Asmussen (1987, §XII.5) for ruin probabilities. For $x > 0$, let

$$T_x = \inf\{n \geq 1 : S_n > x\},$$

and define $\tilde{T}_x$ from $\tilde{S}_n$ analogously. Since S_n has negative drift, T_x may be infinite, but $\tilde{T}_x$ is finite with probability one, for all $x > 0$. Using the equality in distribution of Y and M, the definition of T_x and then Wald's likelihood ratio identity (see §XII.4 of Asmussen 1987 or page 13 of Siegmund) we have

$$\begin{aligned} P(Y > x) &= P(M > x) \\ &= P(T_x < \infty) \\ &= \mathsf{E}[\exp(-\gamma \tilde{S}_{\tilde{T}_x})] \\ &= e^{-\gamma x}\mathsf{E}[\exp\{-\gamma(\tilde{S}_{\tilde{T}_x} - x)\}]. \end{aligned} \tag{41}$$

An application of the renewal theorem shows that

$$C \stackrel{\Delta}{=} \lim_{x\to\infty} \mathsf{E}[\exp\{-\gamma(\tilde{S}_{\tilde{T}_x} - x)\}], \tag{42}$$

exists, the limit taken through integer x for discrete F_D. Thus, $P(Y > x) \sim C \exp(-\gamma x)$ as $x \to \infty$. The expression given for C shows that $C \leq 1$.

We now give

Proof of Theorem 1. Part (i) follows from the analysis just given of $P(Y > x)$ and the definition of $\alpha(s)$. For part (ii), observe that (i) implies that for any $\epsilon > 0$ there is an s_ϵ such that for all $s \geq s_\epsilon$

$$|P(Y > s) - Ce^{-\gamma s}| \leq \epsilon Ce^{-\gamma s}.$$

Consequently,

$$\left| \int_s^\infty P(Y > x)\, dx - \int_s^\infty Ce^{-\gamma x}\, dx \right| \leq \epsilon \int_s^\infty Ce^{-\gamma x}\, dx;$$

i.e.,

$$|b(s) - (C/\gamma)e^{-\gamma s}| \leq \epsilon(C/\gamma)e^{-\gamma s},$$

showing that the ratio of $b(s)$ to $(C/\gamma)e^{-\gamma s}$ differs from 1 by at most ϵ for all sufficiently large s. Part (iii) follows in the same way.

For part (iv), we use (6) to express the expected demands *not* filled in a period, as

$$\begin{aligned} \mathsf{E}[D](1 - \beta(s)) &= \mathsf{E}[(Y + D - c - s)^+; Y \leq c + s] \\ &\quad + \mathsf{E}[D; Y > c + s] \\ &= (\mathsf{E}[(Y + D - c - s)^+] \\ &\quad - \mathsf{E}[(Y + D - c - s)^+; Y > c + s]) \\ &\quad + \mathsf{E}[D; Y > c + s] \\ &= \mathsf{E}[(Y - s)^+] - \mathsf{E}[(Y - c - s)^+] \\ &= \int_s^{s+c} P(Y > x)\, dx, \end{aligned} \tag{43}$$

the third equality using (4). Much as in part (ii), we have

$$\int_s^{s+c} P(Y > x)\, dx \sim (C/\gamma)(1 - e^{-\gamma c})e^{-\gamma s},$$

from which the result follows. □

4.2. Analysis of the Bounds

We now turn to the bounds C_- and C_+, beginning with

Proof of Theorem 2. It follows from (41) that, for all $x > 0$,

$$\begin{aligned} &\inf_{r\geq 0} \mathsf{E}[\exp\{-\gamma(\tilde{S}_{\tilde{T}_r} - r)\}] \\ &\leq e^{\gamma x}P(Y > x) \leq \sup_{r\geq 0} \mathsf{E}[\exp\{-\gamma(\tilde{S}_{\tilde{T}_r} - r)\}]. \end{aligned} \tag{44}$$

As argued in Ross, since the increments of $\{\tilde{S}_n, n \geq 0\}$ have the distribution of $\tilde{X}_1$, a lower bound on the left-most term in (44) is given by

$$\inf_{r\geq 0} \mathsf{E}[\exp\{-\gamma(\tilde{X}_1 - r)\}|\tilde{X}_1 > r]. \tag{45}$$

For each r, an application of Wald's likelihood ratio identity shows that

$$\begin{aligned} &\mathsf{E}[\exp\{-\gamma(\tilde{X}_1 - r)\}|\tilde{X}_1 > r] \\ &= \mathsf{E}[\exp\{-\gamma(\tilde{X}_1 - r)\}; \tilde{X}_1 > r]/P(\tilde{X}_1 > r) \\ &= P(X_1 > r)/\mathsf{E}[\exp\{\gamma(X_1 - r)\}; X_1 > r] \\ &= (\mathsf{E}[\exp\{-\gamma(X_1 - r)\}|X_1 > r])^{-1}. \end{aligned}$$

Writing $D_1 - c$ for X_1 and taking the infimum over r, we find that (45) equals

$$\inf_{r\geq 0}(\mathsf{E}[\exp\{-\gamma(D_1 - c - r)\}|D_1 - c > r])^{-1},$$

which is the same as C_- in (12). The analysis of C_+ works the same way. Thus, from (44) we get

$$C_-e^{-\gamma x} \leq P(Y > x) \leq C_+e^{-\gamma x}, \quad \text{for all } x > 0.$$

Parts (i)–(iv) of Theorem 2 now follow, just as in Theorem 1. □

We turn next to the bounds in Section 2.

Proof of Proposition 4. Part (i) follows from the expression for C given in (42). For part (ii), argue as in Ross to conclude that for NBU demands the infimum in (12) is attained at $r = c$ to get $C_- = e^{-\gamma c}$. For NWU demands, $C_+ = e^{-\gamma c}$ in much the same way. Since Erlang distributions are NBU, the lower bound in part (iii) follows from (ii). For the upper bound, we have from Ross that

$$C_+ = (\lim_{r\to\infty} \mathsf{E}[\exp(\gamma(D_1 - r))|D_1 > r])^{-1}, \tag{46}$$

because Erlang distributions have increasing failure rate (Barlow and Proschan, p. 75). A straightforward calculation shows that if F_D is the $E_m(\mu)$ distribution, then

$$\frac{F_D(x + r) - F_D(r)}{1 - F_D(r)} \to 1 - e^{-\mu x} \quad \text{as } r \to \infty,$$

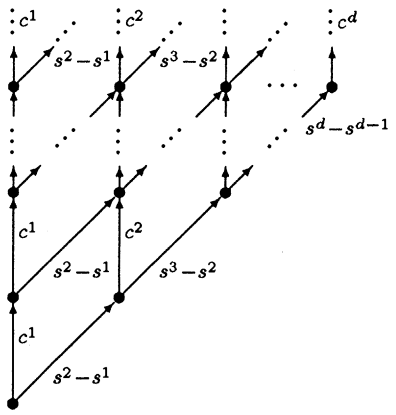

Figure 5. Each vertical arc in column i has length c^i; each diagonal arc from column i to column $i+1$ has length $s^{i+1} - s^i$.

for all $x \geq 0$, indicating that the distribution $D_1 - r$ conditional on $D_1 > r$ converges to an exponential. Moreover, the convergence is monotone, so

$$\lim_{r\to\infty} \mathsf{E}[\exp(\gamma(D_1 - r))|D_1 > r]$$

$$= \int_0^\infty e^{\gamma x}\mu e^{-\mu x}\, dx = \frac{\mu}{\mu - \gamma}.$$

By definition, γ satisfies

$$\left(\frac{\mu}{\mu - \gamma}\right)^m e^{-\gamma c} = 1,$$

so the limit in (46) equals $\exp(\gamma c/m)$, yielding the desired expression for C_+. A similar argument holds for the negative binomial, except that the convergence is to a geometric distribution rather than an exponential.

The upper bound in (iv) follows from (ii) since hyperexponential distributions are NWU; in fact, they have decreasing failure rates (Barlow and Proschan, p. 103), from which it follows (again from Ross) that

$$C_- = (\lim_{r\to\infty} \mathsf{E}[\exp(\gamma(D_1 - r))|D_1 > r])^{-1}.$$

A straightforward calculation shows that the distribution of $D_1 - r$ conditional on $D_1 > r$ converges to an exponential with mean $1/\mu_1$, under our convention that $\mu_1 \leq \mu_2$. The convergence is once again monotone, yielding

$$C_- = \left(\frac{\mu_1}{\mu_1 - \gamma}\right)^{-1}.$$

Part (v) follows from Kingman (1970). □

4.3. Proofs for Multistage Systems

We begin by giving an expression for the distribution of Y_n^1 for all n and thus for the limit Y^1. We detail the case in which $c^d < \min_{i\neq d} c^i$ (i.e., the uppermost node is the unique bottleneck) then discuss the general case.

Consider the graph in Figure 5. As indicated, the vertical arcs in column i all have length c^i, $i = 1, \ldots, d$, and diagonal transitions from column i to column $i+1$ have length $s^{i+1} - s^i$. Let r_n be the length of the shortest n-step path through this graph, starting from the lowest node in column 1; specifically, $r_1 = \min(c^1, s^2 - s^1)$, $r_2 = \min(2c^1, c^1 + s^2 - s^1, s^2 - s^1 + c^2, s^3 - s^1)$, and for general n,

$$r_n = \min(k_1 c^1 + \delta_{k_2}(s^2 - s^1) + k_2 c^2 + \cdots + \delta_{k_d}(s^d - s^{d-1}) + k_d c^d),$$

where $\delta_{k_i} \in \{0, 1\}$, $\delta_{k_2} \geq \delta_{k_3} \geq \cdots \geq \delta_{k_d}$, $\delta_{k_i} = 0 \Rightarrow k_i = 0$, and the minimum is over all nonnegative integers $(k_1, \ldots, k_d)$ with

$$k_1 + \delta_{k_2} + k_2 + \cdots + \delta_{k_d} + k_d = n.$$

Since c^d is the smallest of the capacities, there exists an n_* such that for all $n \geq n_*$, the minimum is attained by setting $k_i = 0$ for $i < d$, $\delta_{k_i} = 1$ for $i = 2, \ldots, d$, and $k_d = n - (d-1)$; that is, the shortest path eventually corresponds to moving diagonally $(d-1)$ steps to the last column and then moving vertically up that column for $(n - d + 1)$ steps. This means that $r_n = (s^d - s^1) + (n - d - 1)c^d$ for all $n \geq n_*$ and hence that

$$n \geq n_* \Rightarrow r_n - nc^d = \eta. \tag{47}$$

The constant η is thus the limit of the sequence $r_n - nc^*$. Upper and lower bounds on this sequence are given by

$$\eta_+ = \max_{n\geq 0}\{r_n - nc^*\} \quad \text{and} \quad \eta_- = \min_{n\geq 0}\{r_n - nc^*\}. \tag{48}$$

We now have

Lemma 2. *Suppose $Y_0^i = 0$, for all $i = 1, \ldots, d$. Then*

$$Y_n^1 \overset{d}{=} \max_{1\leq j\leq n}\left[\sum_{i=1}^{j} D_i - r_j\right]^+;$$

consequently,

$$Y^1 \overset{d}{=} \max_{j\geq 1}\left[\sum_{i=1}^{j} D_i - r_j\right]^+.$$

Proof. From the recursions (28)–(29), it follows that

$$\begin{aligned} Y_1^1 &= \max\{0, D_1 - c^1, D_1 - (s^2 - s^1)\} \\ &= \max\{0, D_1 - r_1\}, \\ Y_2^1 &= \max\{0, D_2 - c^1, D_2 + D_1 - 2c^1, D_2 + D_1 - c^1 \\ &\qquad - (s^2 - s^1), D_2 + D_1 - c^2 - (s^2 - s^1)D_2 \\ &\qquad \cdot - (s^2 - s^1), D_1 + D_2 - (s^3 - s^1)\} \\ &= \max\{0, D_2 - r_1, D_1 + D_2 - r_2\}; \end{aligned}$$

proceeding by induction shows that

$$Y_n^1 = \max\{0, D_n - r_1, D_n + D_{n-1} - r_2, \ldots, D_n + \cdots + D_1 - r_n\}.$$

Since the demands are i.i.d., this has the same distribution as

$$\max\{0, D_1 - r_1, D_1 + D_2 - r_2, \ldots, D_1 + \cdots + D_n - r_n\},$$

which is the first assertion of the lemma. Letting n increase yields the second assertion. □

With this we have

Proof of Theorem 3. Let $r_0 = 0$ and let empty sums be zero so that in Lemma 2 we may omit the positive-part operator by taking the maximum over $j \geq 0$. Writing X_i for $D_i - c^d$, we then have, for any $x > 0$,

$$P(Y^1 > x) = P\left(\max_{n \geq 0} \left\{ \sum_{i=1}^{n} D_i - r_n \right\} > x \right)$$

$$= P\left(\max_{n \geq 0} \left\{ \sum_{i=1}^{n} X_i + [nc^d - r_n] \right\} > x \right)$$

$$= P(T < \infty), \quad \text{where}$$

$$T = \inf\left\{ n \geq 1 \colon \sum_{i=1}^{n} X_i > x - [nc^d - r_n] \right\}.$$

From the distribution of X_1 define a conjugate random walk $\{\tilde{S}_n, n \geq 0\}$ with increment distribution defined from γ just as in (38). Define $\tilde{T}$ from $\tilde{X}_i$ just as T is defined from X_i. Then, by Wald's likelihood ratio identity,

$$P(T < \infty) = \mathsf{E}[\exp(-\gamma \tilde{S}_{\tilde{T}})] = e^{-\gamma x} \mathsf{E}[\exp(-\gamma\{\tilde{S}_{\tilde{T}} - x\})].$$

It remains to evaluate the limit of the expectation on the right as x increases.

Define another stopping time

$$\tilde{T}' = \inf\{n \geq 1 \colon \tilde{S}_n > x + \eta\},$$

and notice that it follows from (47) that $\tilde{T} = \tilde{T}'$ on the event $\{\tilde{T} \geq n_*, \tilde{T}' \geq n_*\}$. Hence, we may write

$$\mathsf{E}[\exp(-\gamma\{\tilde{S}_{\tilde{T}} - x\})]$$

$$= \mathsf{E}[\exp(-\gamma\{\tilde{S}_{\tilde{T}'} - x\}); \min\{\tilde{T}, \tilde{T}'\} \geq n_*]$$

$$+ \mathsf{E}[\exp(-\gamma\{\tilde{S}_{\tilde{T}} - x)\}; \min\{\tilde{T}, \tilde{T}'\} < n_*]. \qquad (49)$$

We analyze the two terms on the right separately. For the first term, we have

$$\mathsf{E}[\exp(-\gamma\{\tilde{S}_{\tilde{T}'} - x\}); \min\{\tilde{T}, \tilde{T}'\} \geq n_*]$$

$$= e^{-\gamma\eta} \mathsf{E}[\exp(-\gamma\{\tilde{S}_{\tilde{T}'} - [x + \eta]\}); \min\{\tilde{T}, \tilde{T}'\} \geq n_*].$$

As $x \to \infty$, both $\tilde{T}$ and $\tilde{T}'$ increase to infinity with probability one; so (applying Theorem 4.4.6 of Chung 1974)

$$\lim_{x \to \infty} \mathsf{E}[\exp(-\gamma\{\tilde{S}_{\tilde{T}'} - [x + \eta]\}); \min\{\tilde{T}, \tilde{T}'\} \geq n_*]$$

$$= \lim_{x \to \infty} \mathsf{E}[\exp(-\gamma\{\tilde{S}_{\tilde{T}'} - [x + \eta]\})] = C,$$

where C is as defined in (42) for a single-stage system with capacity c^d. We have therefore shown that the first term on the right in (49) converges to C; it remains to show that the second term vanishes as x increases. Observe that

$$\mathsf{E}[\exp(-\gamma\{\tilde{S}_{\tilde{T}} - x\}); \min\{\tilde{T}, \tilde{T}'\} < n_*]$$

$$\leq \max_{i < n_*} \exp[-\gamma(r_i - ic^d)] P(\min\{\tilde{T}, \tilde{T}'\} < n_*).$$

As x increases, $\min\{\tilde{T}, \tilde{T}'\} \to \infty$, a.s., so $P(\min\{\tilde{T}, \tilde{T}'\} < n_*) \to 0$.

Dropping the assumption that stage d is the unique bottleneck requires minor modification of the argument. Suppose that an arbitrary node i^* has capacity strictly less than all other nodes. For sufficiently large n, the shortest n-step path through the graph in Figure 6 consists of $(i^* - 1)$ diagonal steps to column i^*, followed by nearly n vertical steps, and possibly followed by up to $(d - i^*)$ diagonal steps. Whether or not these additional diagonal steps are included depends on the relative values of the diagonal arc lengths $s^{i^*+1} - s^{i^*}, \ldots, s^d - s^{d-1}$ and the vertical arc length c^*. In any case, for sufficiently large n, we have $r_n - nc^* = \eta$, just as in (47), but now with η as in (31). If there are multiple nodes with capacity c^*, then eventually $r_n - nc^* = \eta$ with η as in (34). The rest of the argument is the same as before. □

Corollaries 3 and 4 are proved in exactly the same way as their counterparts for single-stage systems give in Section 1.2. Theorem 4 follows from Theorem 3 because stage k evolves in the same way as the lowest stage in a system consisting solely of stages $k, k + 1, \ldots, d$.

ACKNOWLEDGMENT

The author is supported, in part, by the National Science Foundation through grants MSS-9216490 and DMI-9457189. Thanks go to the referees for helpful comments.

REFERENCES

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

ASMUSSEN, S. 1993. *Fundamentals of Ruin Probability Theory*. Book Manuscript, Aalborg University, Denmark.

BARLOW, R. AND F. PROSCHAN. 1975. *Statistical Theory of Reliability and Life Testing*. Holt, Reinhart and Winston, New York.

BRATLEY, P., B. L. FOX, AND L. SCHRAGE. 1983. *A Guide to Simulation*. Springer, New York.

CHANG, C. S. 1994. Stability, Queue Length and Delay of Deterministic and Stochastic Queueing Networks. *IEEE Trans. Automatic Control,* **39,** 913–931.

CHUNG, K. L. 1974. *A Course in Probability Theory*. Second Edition, Academic Press, New York.

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

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, Volume 2*. Second Edition, Wiley, New York.

GLASSERMAN, P. AND S. TAYUR. 1995. Sensitivity Analysis for Base-Stock Levels in Multi-Echelon Production-Inventory Systems. *Management Science* **41,** 263–281.

GLASSERMAN, P. AND S. TAYUR. 1994. The Stability of a Capacitated, Multi-Echelon Production-Inventory System Under a Base-Stock Policy. *Opns. Res.* **42,** 913–925.

GLASSERMAN, P. AND S. TAYUR. 1996. A Simple Approximation for a Multistage Capacitated Production-Inventory System. *Naval Res. Logistics* **43,** 41–58.

KINGMAN, J.F.C. 1970. Inequalities in the Theory of Queues. *J.R. Statist. Soc.* **B 32,** 102–110.

LANGENHOFF, L.J.G. AND W.H.M. ZIJM. 1992. An Analytical Theory of Multi-Stage Production/Distribution Systems. *Statistica Neerlandica,* **44,** 149–174.

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

NEUTS, M.F. 1986. The Caudal Characteristic Curve of Queues. *Adv. Appl. Prob.* **18,** 221–254.

PRABHU, U. 1965. *Queues and Inventories*. Wiley, New York.

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

RAMASWAMI, V. AND D.M. LUCANTONI. 1985. Stationary Waiting Time Distributions in Queues with Phase-Type Service and Quasi-Birth-and-Death Processes. *Stochastic Models,* **1,** 125–136.

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

ROSS, S.M. 1974. Bounds on the Delay Distribution in GI/G/1 Queues. *J. Appl. Prob.* **11,** 417–421.

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

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

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

WHITT, W. 1993. Tail Probabilities with Statistical Multiplexing and Effective Bandwidths in Multi-class Queues. *Telecommunications Systems* **2,** 71–107.