CHAPTER 61

Production-Inventory Systems

DAVID D. YAO
Columbia University

Production-inventory systems are one of the most established subjects in industrial engineering. The focus is on studying inventory dynamics, with inventory viewed as a buffer between supply (production/replenishment) and customer demand. Hence the emphasis is really more on inventory than on production, the latter being the primary focus of other IE subjects, such as scheduling and production planning.

Classical textbook models of production-inventory systems focus on the issue of optimality, in particular, the optimality of simple policies that are characterized by reorder points and order sizes. More recent studies have shifted the emphasis to the inventory–service trade-off (e.g., Cheng et al. 2000; Ettl et al. 2000; Glasserman and Wang 1998; Li 1992; Zipkin 2000) and to topics that are more relevant to industrial applications, such as kanban (e.g., Buzacott and Shanthikumar 1993; Glasserman and Yao 1994b, 1996) and supply chain management (e.g., Cheng et al. 2000; Ettl et al. 2000; Lee and Billington 1986). These topics are the emphasis of this chapter as well.

We start with an overview in Section 1 of several classical models, widely available in textbooks (e.g., Nahmias 1997; Silver et al. 1998). The rest of the chapter has two parts. The first part, Sections 2, 8, and 9, relates several familiar inventory control mechanisms—base-stock control, kanban control, and assemble-to-order systems—to queueing models, emphasizing steady-state results. The theme of the second part, Sections 3, 4, 5, 6, and 7, is nonstationarity, and we use a set of DRP (distribution resource planning)-like recursions as the basic model for inventory dynamics. Finally, in Section 10, the two parts converge into a unified, decomposition-based approach for modeling a supply network. Brief bibliographical notes are given in Section 11.

Handbook of Industrial Engineering: Technology and Operations Management, Third Edition.
Edited by Gavriel Salvendy

1. OVERVIEW OF CLASSICAL MODELS

1.1. The EOQ Model

EOQ stands for economic order quantity. As its name suggests, the EOQ model emphasizes the role of inventory in achieving economies of scale—produce or replenish in batches, rather than single units. Demand is assumed to be deterministic, with rate λ. There is a fixed setup cost of C dollars for placing each replenishment order. Order lead times are zero, which, along with deterministic demand, implies that no order need be placed until the inventory level drops down to zero. The inventory holding cost is charged at a rate of h, that is, for each dollar of inventory that is kept for one unit of time, the charge is h dollars.

Suppose the (batch) size of each order is Q, the decision variable. Then the inventory level starts from the highest point of Q, immediately after an order is placed (and then received instantaneously), and then goes down to zero—depleted by demand, at rate λ. This cycle then repeats itself over the entire time horizon. Hence the average inventory level is $Q/2$ over each cycle and hence over the entire horizon as well.

The total cost—setup cost plus holding cost—per time unit is $\lambda C/Q + hQ/2$, where λ/Q is the number of orders placed per time unit. Note that we have ignored the variable cost, the cost for purchasing the units, since this term is independent of the decision variable Q. It is equal to $c\lambda$ per time unit, with c being the purchasing cost per unit. The Q that minimizes this cost objective is Q : $= \sqrt{2\lambda C/h}$, which is easily derived from setting the derivative of the cost objective (with respect to Q) to zero. The square-root order quantity is often referred to as EOQ as well.

1.2. The Newsvendor Model

This is a single-period problem. Demand, D, is random, with a distribution function $F(x)$ that is known at the beginning of the period. The actual realization of the demand will not be known until the end of the period. The problem is to decide the order quantity Q at the beginning of the period, under the following cost assumptions: Each unit of demand supplied earns a profit (selling price minus cost) of p, each unit of unmet demand incurs a penalty of π, and each surplus (i.e., unsold) unit at the end of the period carries a net loss of ℓ (i.e., cost minus any salvage value). The objective is to maximize the expected net profit:

$$\max_Q \, [p\mathsf{E}(D \wedge Q) - \pi\mathsf{E}(D - Q)^+ - \ell\mathsf{E}(Q - D)^+]$$

where $\wedge$ denotes the min operator and $[x]^+ := \max\{x, 0\}$. The above objective simplifies to

$$\max_Q \, [(p + \pi)Q - (p + \pi + \ell)\mathsf{E}(Q - D)^+]$$

making use of the following identities:

$$D \wedge Q = Q - (Q - D)^+, \quad \text{and} \quad (D - Q)^+ = D - Q + (Q - D)^+$$

and ignoring the term $\pi\mathsf{E}(D)$, which is independent of Q. Hence, setting the derivative of the objective function (with respect to Q) to zero yields

$$F(Q^*) = \frac{p + \pi}{p + \pi + \ell}$$

noticing that

$$\begin{aligned}
&\frac{d}{dQ}\mathsf{E}(Q - D)^+ \\
&= \frac{d}{dQ}\int_0^Q (Q - x)dF(x) \\
&= \frac{d}{dQ}[QF(Q) - \int_0^Q xdF(x) \\
&= F(Q) + QF'(Q) - QF'(Q) \\
&= F(Q)
\end{aligned}$$

1.3. Deterministic Multiperiod Model

Let $k = 1, \ldots, n$ index the periods; let d_k denote the demand in period k. Demand in all n periods is deterministically known.

A production or replenishment can be ordered at the beginning of each period; the decision is the lot size—how many subsequent periods' demand this order should cover. For instance,we can decide that an order placed at the beginning of period i to have a lot size of $d_i + \ldots + d_j$, so as to cover the demand in period i through period j, for some $j > i$. That is, we run a single production at the beginning of period i, to supply the demand in all subsequent periods up to j. The cost associated with this decision is:

$$c_{ij} = C + h[d_{i+1} + 2d_{i+2} + \ldots + (j - i)d_j]$$

where C is the fixed setup cost and h is the inventory holding cost per unit of inventory, per period—applied to period-end inventory only.

Note that, as in the EOQ model, there is zero lead time and the variable cost is not included in c_{ij}, as it is independent of the lot-sizing decisions. Also, no back order is allowed, so that each period's demand must be supplied either by production at the beginning of that period or by inventory carried over from the previous periods.

This model can be solved by dynamic programming (DP). Let V_k be the optimal cost-to-go in period k, that is, the total cost to supply the demand in periods k through period n, following the optimal lot-sizing decisions. Then V_1 is the desired solution, which we can derive through the DP recursion as follows.

Clearly, in period n, where there is only a single period left, the only possible solution is to place an order of d_n units to supply the demand; hence, $V_n = C$. Recursively, for period $i = n - 1, \ldots, 2, 1$, we have

$$V_i = \min_{j \geq i} \{c_{ij} + V_{j+1}\}$$

following the DP principle of optimality. That is, the decision in period i is to pick a future period j—that is, to make a lot-sizing decision on the production so as to cover the demand in period i through period j—so as to minimize the right-hand side of the above recursion. Once j is selected, the remaining problem (i.e., the one that starts from period $j + 1$) was already solved in the earlier stages of the DP recursion, that is, to follow the optimal decision embodied in V_{j+1}.

1.4. Other Standard Models

When demand is random, a standard model is to assume that the demand stream follows a renewal process, that is, $\{D_n\}$ are independent and identically distributed (i.i.d.) random variables, where D_n denotes demand in period n. Suppose each production/replenishment order requires a lead time, denoted L, which can be constant or random.

In this context, in addition to the order size, we need a second parameter, the *reorder point*. Associated with the latter is the notion of *inventory position,* defined as the on-hand inventory minus any back orders plus any outstanding orders—orders that have been placed but have not yet arrived due to the lead time. Hence, whenever the inventory position falls below the reorder point, an order is placed. Note that since demand now fluctuates randomly and there is a lead time between placing and receiving an order, it is not desirable, in general, to order only when the inventory drops down to zero, as in the case of the EOQ model.

The above describes exactly how the so-called (Q, R) model works, with R being the reorder point and Q the order size. One way to set these parameters is to let Q take the EOQ value and let R be determined by service requirements. For instance, set the value of R sufficiently large so as to ensure that the stockout probability (i.e., the probability that the on-hand inventory is zero upon a demand arrival) is limited to, say, no more than 5%, or the fill rate (the proportion of demand that is filled from on-hand inventory) is at least 95%. (Note that in general the no-stockout probability is not equal to the fill rate.) It is also possible to set up a cost objective and then optimize it to derive the best Q and R values jointly.

Closely related, but more general, is the (S, s) model, which works as follows: whenever the inventory position falls below the lower threshold, s, place an order to bring the inventory position to the upper threshold S. When demand comes in batches, typically there will be an undershoot (of random size) when the inventory position falls below s. Hence, in the (S, s) model, the order size is random, in contrast to the constant order size in the (Q, R) model.

2. BASE-STOCK CONTROL

A widely studied inventory control mechanism is *base-stock control*, which works as follows: whenever the inventory position drops below R, a constant parameter, place an order to bring the inventory position back to R. This way, every demand will trigger the placement of an order, such that the overall inventory position is always maintained at a constant level, R, which is called the *base-stock level*.

In the case when demand always comes in unit size, then the base-stock control is a special case of the (Q, R) model, with $Q = 1$. Hence, the base-stock control is sometimes also referred to as the one-for-one replenishment rule.

2.1. The Inventory-Queue Model

The base-stock control model relates to a queueing model as follows. Consider the case of demand in single units mentioned above. (Batch demand will be discussed below in Section 2.3.) Let each unit of demand correspond to a job arrival to the queue. Each unit of demand, on arrival, is supplied by the on-hand inventory or, in the stock-out situation, joins a back-order queue. Regardless, however, a replenishment/production order is triggered.

Let the outstanding orders correspond to the jobs in the queueing system. In the inventory system, each of these orders will materialize (or ''arrive''—become available to supply demand) after a lead time. This corresponds to a queueing system with an infinite number of servers, so that any job will be served immediately on arrival. Hence, the overall cycle time of each job in the system is simply its service time, which corresponds to the lead time of orders in the inventory system.

Let N be the number of jobs in the queueing system in steady state. This, as explained above, represents the number of outstanding orders. If $N < R$, then there are $R - N$ units of on-hand inventory. If $N > R$, then we know there are $N - R$ units of back orders: R units of demand have been supplied with on-hand inventory, while the remaining $N - R$ units are back ordered. Hence, with I and B denoting the on-hand inventory and the back orders, respectively, and with $[x]^+$ denoting $\max\{x, 0\}$, we have

$$I = [R - N]^+, \quad B = [N - R]^+ \tag{1}$$

Note that the above implies

$$I - B = R - N$$

and $I \cdot B = 0$, that is, I and B cannot both be positive.

Suppose demand follows a Poisson process with rate λ, and suppose the lead time is L, the time it takes to process and finish an order. Then the queueing system in question is an $M/G/\infty$ model (see, e.g., Wolff 1989), and it is known that N follows a Poisson distribution with mean $\rho := \lambda \mathsf{E}(L)$. That is,

$$\mathsf{P}[N = n] = \frac{\rho^n}{n!} e^{-\rho}, \quad n = 0, 1, 2, \ldots$$

Combining the above with (1), we can derive the distributions of the on-hand inventory and back orders:

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = 1 - \sum_{n=0}^{R-1} \frac{\rho^n}{n!} e^{-\rho}$$

$$\mathsf{P}[I = n] = \mathsf{P}[N = R - n] = \frac{\rho^{R-n}}{(R - n)!} e^{-\rho}, \quad n = 1, \ldots, R$$

and

$$\mathsf{P}[B = 0] = \mathsf{P}[N \leq R] = \sum_{n=0}^{R} \frac{\rho^n}{\mathrm{n}!} e^{-\rho},$$

$$\mathsf{P}[B = n] = \mathsf{P}[N = R + n] = \frac{\rho^{R+n}}{(R + n)!} e^{-\rho}, \quad n = 1, 2, \ldots .$$

Next, suppose that due to limited production capacity, replenishment/production orders may have to wait in queue before they can be processed. In other words, in addition to the lead time (i.e., processing time or "service time"), there is also queueing time. In this case, a queueing model with a finite number of servers is more appropriate. Suppose we use the $M/M/1$ model for the corresponding queueing system. Then, N follows a geometric distribution:

$$\mathsf{P}[N = n] = \rho^n(1 - \rho), \qquad n = 0, 1, 2, \ldots$$

The distributions of I and B in the case are:

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = \rho^R,$$
$$\mathsf{P}[I = n] = \mathsf{P}[N = R - n] = \rho^{R-n}(1 - \rho), \quad n = 1, \ldots, R$$

and

$$\mathsf{P}[B = 0] = \mathsf{P}[N \leq R] = 1 - \rho^{R+1}$$
$$\mathsf{P}[B = n] = \mathsf{P}\ [N = R + n] = \rho^{R+n}(1 - \rho), \quad n = 1, 2, \ldots$$

Note from (1) that in either case, while the on-hand inventory is limited to $I \leq R$, the base-stock level, there is no upper limit on the number of back orders, B.

2.2. Normal Approximations

We can also approximate the above Poisson distribution by a normal distribution, and write

$$N = \rho + Z\sqrt{\rho}$$

where Z denotes the standard normal variate, with density function $\phi(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ and distribution function $\Phi(x)$. Below we shall also use $\overline{\Phi}(x) := 1 - \Phi(x)$. We can also write

$$R = \rho + k\sqrt{\rho}$$

where k is known as the "safety factor" in the inventory literature.

In fact, the normal approximation applies more generally; it does not have to be restricted to N following a Poisson distribution as in the inventory-queue models of Section 2.1. We can start with writing

$$N = \mu + \sigma Z \tag{2}$$

where $\mu := \mathsf{E}(N)$ and $\sigma := \mathsf{sd}(N)$; and then write

$$R = \mu + k\sigma \tag{3}$$

accordingly.

This way, we have

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = \mathsf{P}[Z \geq k] = \overline{\Phi}(k) \tag{4}$$

which is the stockout probability.

To derive $\mathsf{E}(I)$ and $\mathsf{E}(B)$, we first define the following two functions:

$$G(x) := \mathsf{E}[Z - x]^+ = \int_x^\infty (z - x)\phi(z)\,dz \tag{5}$$

and

$$H(x) := \mathsf{E}(x - Z)^+ = x + G(x) = \mathsf{E}[x - Z + (Z - x)^+] = x + \mathsf{E}(Z - x)^+ = x + G(x) \tag{6}$$

We summarize below the properties of the functions $G(x)$ and $H(x)$. These properties are easily verified from first principle.

1. $G(x) = \phi(x) - x\overline{\Phi}(x)$ (note that $\phi'(x) = -x\phi(x)$); and $H(x) = \phi(x) + x\Phi(x)$. $G'(x) = -\overline{\Phi}(x)$, $H'(x) = \Phi(x)$, and $G''(x) = H''(x) = \phi(x)$.
2. For all $x \in (-\infty, +\infty)$, $G(x) \geq 0$ and $H(x) \geq 0$; $G(x)$ is decreasing and $H(x)$ increasing, and

both are convex. (Throughout, we use "increasing" and "decreasing" in the nonstrict sense.) Note in particular that $(x)^+$ is increasing and convex in x.

3. For $x >> 1$, $G(x) \cong 0$ and $G(-x) \cong x$, whereas $H(x) \cong x$ and $H(-x) \cong 0$. In fact, these approximations work very well for $x \geq 2$; for instance, $G(-2) = 2.0085$ and $G(2) = 0.0085$.

Hence, incorporating (2) and (3) with (1) and making use of the G and H functions, we can derive:

$$\mathsf{E}(I) = \sigma\mathsf{E}[k - Z]^+ = \sigma H(k), \quad \text{and} \quad \mathsf{E}(B) = \sigma\mathsf{E}[Z - k]^+ = \sigma G(k) \tag{7}$$

2.3. Demand and Demand over Lead Time

Let $D(t)$ be the demand in period t, $t = 1, 2, \ldots$ Suppose demand (per period) over time is independent and identically distributed. Let L denote the lead time to fill each replenishment order. The number of outstanding orders, as explained in Section 2.1, is equal to the number of jobs, N, in an infinite-server queueing system. In particular, if the per-period demand follows a Poisson distribution, then N also follows a Poisson distribution with mean $\mathsf{E}(N) = \mathsf{E}(D) \cdot \mathsf{E}(L)$ ($= \rho$ in Section 2.1; here D denotes the generic per-period demand). Since N follows a Poisson distribution, we know $\text{Var}(N) = \mathsf{E}(N)$.

On the other hand, the total demand over the lead time,

$$D(1, L) := D(1) + \mathrm{D}(2) + \ldots + D(L)$$

has a mean

$$\mathsf{E}D(1, L) = \mathsf{E}(D) \cdot \mathsf{E}(L)$$

and a variance

$$\mathsf{Var}D(1, L) = \mathsf{Var}(D) \cdot \mathsf{E}(L) + \mathsf{E}^2(D) \cdot \mathsf{Var}(L)$$

Hence, while the lead time demand has the same mean as $\mathsf{E}(N)$, its variance is different, unless the lead time is deterministic and D follows a Poisson distribution. This distinction is important, since it is N, *not* demand over lead time, that determines the inventory performance, in terms of on-hand inventory (I) and back orders (B) via the relations in (1).

In applications, it is often more convenient to model the demand as a *batch* Poisson arrival process. Let λ be the demand arrival rate, as in Section 2.1, and let X denote the batch size associated with each arrived demand. Assume the batches are i.i.d. and independent of the demand arrival process, with $\mathsf{E}(X) \geq 1$. Then the per-period (or per-time unit) demand has the following mean and variance:

$$\mathsf{E}(D) = \lambda\mathsf{E}(X), \quad \text{and} \quad \mathsf{Var}(D) = \lambda\mathsf{E}(X^2)$$

The main advantage of this demand model is that it has (at least) three parameters: the arrival rate λ and the first two moments of the batch size X; whereas the Poisson demand model has only one parameter (the mean, or arrival rate).

Accordingly, the queue model $M/G/\infty$ of Section 2.1) now becomes $M^X/G/\infty$. Following the queueing result in (Liu et al. 1990), the mean and the variance of N are as follows:

$$\mathsf{E}(N) = \lambda\mathsf{E}(X)\mathsf{E}(L) = \mathsf{E}(D)\mathsf{E}(L) \tag{8}$$

$$\begin{aligned}
\mathsf{Var}(N) &= \mathsf{E}(N) + \lambda[\mathsf{E}(X^2) - \mathsf{E}(X)] \int_0^\infty \overline{F}_L^2(y)\, dy \\
&\leq \mathsf{E}(N) + \lambda[\mathsf{E}(X^2) - E(X)] \int_0^\infty \overline{F}_L(y)\, dy \\
&= \mathsf{E}(N) + \lambda[\mathsf{E}(X^2) - \mathsf{E}(X)]\mathsf{E}(L) \\
&= \lambda\mathsf{E}(X^2)\mathsf{E}(L) \\
&= \mathsf{Var}(D)\mathsf{E}(L)
\end{aligned} \tag{9}$$

where $F_L(y) = 1 - \overline{F}_L(y)$ denotes the distribution function of the lead time L, the inequality is due to $\overline{F}_L^2(y) \leq \overline{F}_L(y)$, and $\mathsf{E}(X^2) \geq \mathsf{E}^2(X) \geq \mathsf{E}(X)$ (since $\mathsf{E}(X) \geq 1$). Note again that the variance of N in (9) is different from the variance of lead time demand. In fact, the $\mathsf{Var}(N)$ is even smaller than the first of the two terms of the variance of lead time demand.

The generating function of N is known for the $M/G/\infty$ model (refer to Liu et al. 1990), from which the distribution of N can be derived, at least in principle. It is, however, much simpler to derive the mean and the variance of N and then approximate it by a normal distribution, as in Section 2.2.

The queueing quantity N, by definition, is nonnegative. Hence, to approximate it with a normal distribution is not always appropriate. Sometimes an adjustment to the normal distribution is needed. The same applies if we model demand D using a normal distribution.

Specifically, instead of writing $N = \mu + \sigma Z$, where Z is the standard normal variate, we should have $\tilde{N} = [\mu + \sigma Z]^+$. The mean of $\tilde{N}$ follows from (5):

$$\mathsf{E}[\tilde{N}] = \sigma \mathsf{E}\left[Z - \left(-\frac{\mu}{\sigma}\right)\right]^+ = \sigma G\left(-\frac{\mu}{\sigma}\right) \tag{10}$$

To derive the variance of $\tilde{N}$, note the following:

$$\begin{aligned}\mathsf{E}\{[(Z-x)^+]^2\} &= \int_x^\infty (z-x)^2\phi(z)\,dz \\ &= x\phi(x) + \overline{\Phi}(x) - 2x\phi(x) + x^2\overline{\Phi}(x) \\ &= \overline{\Phi}(x) - xG(x)\end{aligned}$$

where the last equation makes use of (5). Hence,

$$\begin{aligned}\mathsf{Var}[\tilde{N}] &= \sigma^2 \mathsf{Var}\left\{\left[Z - \left(-\frac{\mu}{\sigma}\right)\right]^+\right\} \\ &= \sigma^2 \mathsf{E}\left\{\left[\left(Z - \left(-\frac{\mu}{\sigma}\right)\right)^+\right]^2\right\} - [\mathsf{E}(\tilde{N})]^2 \\ &= \sigma^2\left[\overline{\Phi}\left(-\frac{\mu}{\sigma}\right) + \frac{\mu}{\sigma} G\left(-\frac{\mu}{\sigma}\right) - G^2\left(-\frac{\mu}{\sigma}\right)\right]\end{aligned} \tag{11}$$

For moderately large x (say, $x \geq 2$), from property (c) of the G function in Section 2.2, we have $G(-x) \cong x$, and hence

$$\mathsf{E}[\tilde{N}] \cong \mathsf{E}[N], \quad \mathsf{Var}[\tilde{N}] \cong \mathsf{Var}[N]$$

from (10) and (11). Therefore, the above adjustment is *not* needed when the coefficient of variation, σ/μ, is relatively small, say, $\sigma/\mu < 0.5$.

3. NONSTATIONARY DEMAND: THE DRP FRAMEWORK

DRP stands for distribution resource planning, a class of commercial software tools that are widely used in industry for managing a firm's production/inventory/distribution systems. When demand is nonstationary, the dynamics of inventory are best captured by a set of recursive equations, which are closely related to the logic built into DRP software. Hence, we present next an overview of DRP, or more precisely, a formal abstraction of standard DRP procedures, the latter being widely available in professional references (e.g., Martin 1990; and Stenger 1994).

Time is discrete, indexed by $t = 1, 2, \ldots, n$, and referred to as periods. Suppose we are currently at the beginning of period 1 (or, the end of period 0). We are interested in the following quantities for all future periods, $t = 1, 2, \ldots, n$:

I_t = the on-hand inventory at the end of time period t
B_t = the back-ordered demand at the end of time period t
A_t = the required quantity of product needed at the beginning of period t
Q_t = the constrained (or feasible) quantity of product needed at the beginning of period t
$\tilde{Q}_t$ = the recommended order quantity at the beginning of period t

The distinction between the requirements, A_t, and the constrained order quantity, Q_t, is important. Whereas A_t represents the quantity that is needed at the beginning of period t, Q_t reflects what is feasible, taking into account lead time constraints and order quantity restrictions. For example, if $A_t = 40$ and the maximum order quantity is 30, then Q_t would be set to 30.

The following information is assumed known at the start of time period 1, for all future periods $t = 1, \ldots, n$:

D_t = the demand, in terms of its distribution, in particular $\mathsf{E}[D_t]$ and $\mathsf{sd}[D_t]$
Σ_t = the safety stock requirement, which is the portion of on-hand inventory that should be maintained in the period to protect against demand uncertainty so as to achieve a prescribed service-level requirement
I_0 = the on-hand inventory at the beginning of time period 1
B_0 = the back-ordered demand at the beginning of time period 1

In DRP, there is also the quantity of scheduled receipts, quantities that are in transit and scheduled to arrive at (the start of) each future period. Here we simply ignore these because as they can be easily netted from the inventory/back order of each period or added to Q_t.

At the beginning of each period t, the following sequence of events takes place. First, replenishment (including any scheduled receipt) arrives, that is, Q_t, the constrained order quantity (which relates to A_t and will be specified below). These units are used first to satisfy the back orders, if any, left from the previous period. Next, demand of the period, D_t, is realized and filled, which brings us to the end of the period, when I_t (on-hand inventory) and B_t (back orders) are updated.

Hence, we have the following recursive relation:

$$A_t = [B_{t-1} - I_{t-1} + D_t + \Sigma_t]^+ \tag{12}$$

which says that the replenishment requirement in each period, along with any on-hand inventory from the last period should be able to supply the demand of the current period and any backlog from the last period and still result in a surplus that is equal to the required safety stock for this period.

Next, the constrained order quantity, Q_t, is derived from A_t by applying a set of prespecified order-size restrictions, referred to as *order policies or order rules.* Several commonly used rules are listed below. Note that all of the rules below apply if and only if $A_t > 0$; otherwise, we simply set $Q_t = A_t = 0$, as $0 \le Q_t \le A_t$ by definition.

1. *Lot-for-lot:* $Q_t = A_t$, i.e., there is no restriction on the order quantity.
2. *Min-max:* With $Q_{\min}$ and $Q_{\max}$ being the (given) lower and upper limit on order quantities, $Q_t = \min\{\max(Q_{\min}, A_t), Q_{\max}\}$, becomes the order quantity planned for period t
3. *DOS* (days of supply): Q_t is equal to the (projected) demand over a given number of periods.
4. *EOQ* (economic order quantity): $Q_t = \sqrt{2C\mathsf{E}(D_t)/h}$, the classical EOQ formula (refer to Section 1.1), where C is a fixed cost for placing a replenishment order, and h is the inventory holding cost per period.

Note that all the rules above, with the exception of 1, impose some restrictions on the order quantity. In general, we can assume Q_t relates to A_t (and other parameters) via some prespecified function.

Define the *net* inventory level, Y_t, at the beginning of period t:

$$Y_t = I_{t-1} - B_{t-1}, \quad t = 1, \ldots, n \tag{13}$$

Then clearly,

$$I_t = [Q_t + Y_t - D_t]^+, \quad \text{and} \quad B_t = [D_t - Q_t - Y_t]^+ \tag{14}$$

Finally, suppose the order lead time is L (periods). Then the recommended order quantity, $\tilde{Q}_t$, in period t, corresponds to the calculated (constrained) order quantity $L + t$ periods later:

$$\tilde{Q}_t = \mathrm{Q}_{t+L}, \quad t = 1, \ldots, n - L$$

So the recommended order quantity at the start of period t is whatever the constrained order quantity is for period $t + L$.

4. LOT-FOR-LOT POLICY

Assume that the demand in period t can be expressed as

$$D_t = \mu_t + \sigma_t Z_t$$

where μ_t is the mean demand in period t, σ_t is the standard deviation of demand in period t, and $\{Z_t, t = 1, \ldots, n\}$ are i.i.d. random variables following any distribution with zero mean and unit variance. For ease of discussion, we shall focus on the case in which Z_t follows a standard normal distribution. The results extend readily to more general demand distributions.

Suppose, as before, that the lead time is L and we want to determine the requirement A_t for period t. We need to make this decision at the beginning of period $t - L$ so as to place the order in time for it to arrive in period t.

Position ourselves at the beginning of period $t - L$. What we know are the following: (a) the net inventory left from the previous period, Y_{t-L} and (b) the scheduled arrivals, $A_{t-L}, A_{t-L+1}, \ldots, A_{t-1}$, which had already been decided and ordered. What we do *not* know is the sequence of demand over the lead time, $D_{t-L}, D_{t-L+1}, \ldots, D_t$.

Denote:

$$D(s, t) := D_s + \cdots + D_t, \quad A(s, t) := A_s + \cdots + A_t$$
$$\mu(s, t) := \mathsf{E}[D(s, t)], \quad \sigma(s, t) := \mathsf{sd}[D(s, t)]$$

We claim that the right A_t value should be:

$$A_t = [\mu(t - L, t) + k_t\sigma(t - L, t) - Y_{t-L} - A(t - L, t - 1)]^+. \tag{15}$$

First, suppose $A_t > 0$, i.e., the quantity inside $[\cdot]^+$ on the right hand side above, is positive. Then, at the beginning of period $t - L$, the net inventory is Y_{t-L}, which, along with the scheduled arrivals over the lead time—a total of $A(t - L, t)$, is going to supply all the demands over the lead time—a total of $D(t - L, t)$. Hence, at the end of period t, the expected back order is:

$$\begin{aligned}\mathsf{E}[B_t] &= \mathsf{E}[D(t - L, t) - Y_{t-L} - A(t - L, t)]^+ \\ &= \mathsf{E}[\mu(t - L, t) + Z_t\sigma(t - L, t) - \mu(t - L, t) - k_t\sigma(t - L, t)]^+ \\ &= \sigma(t - L, t)\, \mathsf{E}[Z_t - k_t]^+ \\ &= \sigma(t - L, t)\, G(k_t)\end{aligned} \tag{16}$$

where the second equality follows from (15) since $A_t > 0$ as assumed:

$$Y_{t-L} + A(t - L, t) = \mu(t - L, t) + k_t\sigma(t - L, t)$$

Similarly,

$$\begin{aligned}\mathsf{E}[I_t] &= \mathsf{E}[Y_{t-L} + A(t - L, t) - D(t - L, t)]^+ \\ &= \sigma(t - L, t) + \mathsf{E}[k_t - Z_t]^+ \\ &= \sigma(t - L, t)\, H(k_t)\end{aligned} \tag{17}$$

Next, suppose $A_t = 0$ in (15). This implies

$$Y_{t-L} + A(t - L, t - 1) \geq \mu(t - L, t) + k_t\sigma(t - L, t)$$

Hence,

$$\begin{aligned}\mathsf{E}[B_t] &= \mathsf{E}[\mu(t - L, t) + Z \cdot \sigma(t - L, t) - Y_{t-L} - A(t - L, t - 1)]^+ \\ &= \sigma(t - L, t)\, G(k_t')\end{aligned} \tag{18}$$

with

$$k_t' = \frac{Y_{t-L} + A(t - L, t - 1) - \mu(t - L, t)}{\sigma(t - L, t)} \geq k_t \tag{19}$$

Hence, $\mathsf{E}[B_t]$ in (18) is no greater than $\mathsf{E}[B_t]$ in (16), since $G(\cdot)$ is a decreasing function. (Intuitively, in this case, since the available (net) inventory Y_{t-L} is higher, the effective safety factor k_t' is also higher; hence, the (projected) back order is lower.)

Similarly, when $A_t = 0$, we have

$$\mathsf{E}[I_t] = \sigma(t - L, t)\, H(k_t') \tag{20}$$

with k_t' following (19); and $\mathsf{E}[I_t]$ above is larger than $\mathsf{E}[I_t]$ of (17).

To summarize, when the order rule is lot-for-lot, that is, $Q_t = A_t$ for all t, the required quantity needed at the beginning of period t (A_t), and hence the recommended order quantity at the beginning of period $t - L$ ($\tilde{Q}_{t-L}$), should be

$$\tilde{Q}_{t-L} = Q_t = A_t = [s_{t-L} - Y_{t-L} - A(t-L, t-1)]^+$$

where

$$s_{t-L} = \mu(t-L, t) + k_t\sigma(t-L, t)$$

Note that in this case, DRP effectively follows a base-stock control mechanism, with s_{t-L} being the base-stock level (or, reorder point) for period $t - L$. The estimates for expected on-hand inventory and back orders (for period t) follow (17, 20) and (16, 18), respectively.

Also note that the time index of the reorder point, s_{t-L}, is consistent with the time when the order is placed. On the other hand, the safety factor k_t is indexed by t since it relates more closely to the on-hand inventory and back orders in period t.

We should point out that the expression in (15) that specifies the requirement A_t can also be derived directly from the DRP logic as follows. From (13) and (14), with $Q_t = A_t$, we have

$$\begin{aligned} Y_t &= I_{t-1} - B_{t-1} \\ &= Y_{t-1} + A_{t-1} - D_{t-1} \\ &= Y_{t-2} + A_{t-2} - D_{t-2} + A_{t-1} - D_{t-1} \\ &= \ldots \\ &= Y_{t-L} + A(t-L, t-1) - D(t-L, t-1) \end{aligned} \tag{21}$$

Now, substituting (13) into (12), we can write the latter as:

$$A_t - [D_t + \Sigma_t - Y_t]^+$$

Substituting (21) into the above and treating demand as deterministic, which is what DRP does, we have

$$D_t + D(t-L, t-1) = \mu(t-L, t)$$

and hence recovering (15). In other words, if DRP uses (15) directly to generate the requirements A_t, instead of going through the recursions that involve I_t and B_t, then the estimates on these requirements will be protected from the in accuracies involved in estimating I_t and B_t.

Following the expressions derived above for A_t, B_t, and I_t, the recursion can be carried over to future periods through $Y_{t+1} = I_t - B_t$. Specifically, starting from period 1, we first derive A_{L+1}, followed by B_{L+1} and I_{L+1}, and then Y_{L+1}. Continue this procedure for $t = L + 2, \ldots, 2L$, assuming that $Y_1, Y_2, \ldots, Y_L$ have all been prespecified, i.e., given at beginning of period 1. (Note that nothing in periods 1 through L can be affected by any replenishment decision, at the beginning of period 1, due to the lead time.) In return, $Y_{L+2}, \ldots, Y_{2L}$ are derived. Next, we will derive A_{2L+1}, at the beginning of period $L + 1$. At that point, we will need the value of Y_{L+1}, which has already been derived; and in return, we will derive Y_{2L+1}. The recursion then continues.

5. (*S*, *s*) POLICY

Having analyzed the lot-for-lot rule, we next want to study other rules that impose restrictions on the order quantities. It will become evident below (Section 6) that all these rules can be unified into a dynamic (S, s) control scheme—dynamic in the sense that the control parameters, S and s, in general change over time. In order to analyze the dynamic (S, s) control rule, here we first review the standard (S, s) inventory model under stationary demand and develop some simple approximations for its performance.

The (S, s) policy is known to be optimal in a quite general setting, in terms of minimizing the total costs of placing orders, keeping inventory, and paying penalty of back orders (refer to Clark and Scarf 1960; Scarf 1960). Our focus here, however, is on performance evaluation rather than cost minimization. In particular, we want to derive certain simple approximate formulas that are not only easy to evaluate but also useful in suggesting approximations in the non-stationary (i.e., dynamic) setting. More refined approximations for the stationary (S, s) systems are available in the literature (e.g., Tijms 1994), although it is not clear how they can be adapted to the nonstationary setting.

5.1. Stationary Analysis

Consider an (S, s) inventory model with periodic review. A review takes place at the beginning of each period t, at which time the inventory position is updated and a replenishment decision is made following the (S, s) rule—order up to S if and only if the inventory position (on-hand plus on-order minus back order) falls below s.

Observe that the inventory position, X_t, as defined above, always takes values between s and S. In particular, if an arriving demand brings the on-hand inventory level to *below* s, the inventory position is immediately brought up to S through placing a replenishment order.

Suppose, without loss of generality, that we start with $X_0 = S$. Then, from standard renewal theory (Ross 1996), we know the time between two consecutive replenishment orders forms a regenerative cycle. Given x ($x = s, s + 1, \ldots, S$), let

$$T_x = \min \{t\colon D_1 + \ldots + D_t > S - x\}$$

be the time, in a cycle, for the inventory position to drop from S to below x. In particular, T_s is the cycle length.

Since demands are i.i.d., we can view $T_x - 1$ as the number of renewals by time $S - x$ of a renewal process with interarrival time D_t. (Refer to Ross 1996, Section 3.4, in particular pp. 69–70.) Hence,

$$\mathsf{E}[T_x] = M(S - x) + 1, \quad \mathsf{E}[T_s] = M(S - s) + 1$$

where

$$M(x) = \sum_{n=1}^{\infty} F_{*n}(x)$$

is the renewal function, with $F_{*n}(\cdot)$ denoting the n-fold convolution of the (per-period) demand distribution, $F(\cdot)$. Furthermore, the limiting (stationary) inventory position, X, has the following distribution:

$$\mathsf{P}[X \geq x] = \lim_{t\to\infty} \mathsf{P}[X_t \geq x] = \frac{\mathsf{E}[T_x]}{\mathsf{E}[T_s]} = \frac{1 + M(S - x)}{1 + M(S - s)}$$

for $x \in [s, S]$. Alternatively,

$$\mathsf{P}[X = x] = \frac{M(S - x) - M(S - x - 1)}{1 + M(S - s)}, \quad x = s, \ldots, S - 1$$

$$\mathsf{P}[X = S] = \frac{1}{1 + M(S - s)}$$

The mean of X then follows:

$$\mathsf{E}[X] = \frac{S + sM(S - s) + \sum_{x=1}^{S-s-1} M(x)}{1 + M(S - s)}$$

To characterize the on-hand inventory and the back orders, the key is to observe that by time t all orders that are placed before or at $t - L$ will have arrived. In other words, the on-order quantities included in the inventory position at $t - L$ will all have arrived by t. As before, let $D(t - L, t)$ denote the total demand over the $L + 1$ periods: $t - L, t - L + 1, \ldots, t$. We have

$$I_t = [X_{t-L} - D(t - L, t)]^+ \tag{22}$$

where $[x] + = \max\{x, 0\}$. Note that in (22), $D(t - L, t)$ is independent of X_{t-L}. Similarly, we have

$$B_t = [D(t - L, t) - X_{t-L}]^+ \tag{23}$$

Letting $t \to \infty$, and denoting the limits by omitting the time index, we have

$$I = [X - D(L + 1)]^+, \quad \text{and} \quad B = [D(L + 1) - X]^+ \tag{24}$$

where $D(L + 1)$ denotes a random variable that is equal in distribution to $D(t - L, t)$ and is independent of X.

5.2. Approximations

The renewal function is known to have a linear asymptote:

$$M(x) \sim \frac{x}{\mu} + \frac{m_2}{2\mu^2} - 1 \tag{25}$$

for large x, where $\mu = \mathsf{E}(D)$ and $m_2 = \mathsf{E}(D^2)$ are the first two moments of the (per-period) demand distribution. (More precisely, the difference between $M(x)$ and the linear function on the right hand side above goes to zero when $x \to \infty$).

Since in most applications the (S, s) values are large or moderately large, we can use the linear asymptote in (25) as an approximation for the renewal function. This way, the probability distribution of the inventory position X in Section 5.1 becomes:

$$\mathsf{P}[X = x] = \frac{1}{\dfrac{m_2}{2\mu} + S - s}, \quad x = s, \ldots, S - 1$$

$$\mathsf{P}[X = S] = \frac{\dfrac{m_2}{2\mu}}{\dfrac{m_2}{2\mu} + S - s}$$

where $m_2/2\mu$ is the average "undershoot"—the gap between the inventory position when an order is placed (which is necessarily smaller than s) and the reorder point s. Hence,

$$\mathsf{E}[X] = \frac{(1/2)(S + s - 1)(S - s) + (m_2/2\mu)\, S}{m_2/2\mu + S - s} = \frac{\mu(S - s)(S + s - 1) + m_2 S}{m_2 + 2\mu(S - s)} \tag{26}$$

In some special cases, the above reduces to a uniform distribution over the integers $\{s, \ldots, S\}$ and $\mathsf{E}[X] = (S + s)/2$. For instance, this is the case when the (per-period) demand follows a truncated normal distribution with unit mean and unit coefficient of variation (hence, $\mu = 1$ and $m_2 = 2$). Note that for truncated normal distributions, just as for normal distributions, assuming a unit coefficient of variation loses no generality; hence, unit mean is the only significant assumption here.

Next, making use of (24), along with the normal approximation of $D(L + 1)$, we get the following approximation:

$$\mathsf{E}[B] = \sum_{x=s}^{S} \sigma\sqrt{L + 1}\, G\left(\frac{x - (L + 1)\mu}{\sigma\sqrt{L + 1}}\right) \mathsf{P}[X = x]$$

Substituting the uniform distribution of X into the above, we have

$$\begin{aligned} \mathsf{E}[B] &= \frac{1}{1 + S - s} \sum_{x=s}^{S} \sigma\sqrt{L + 1}\, G\left(\frac{x - (L + 1)\mu}{\sigma\sqrt{L + 1}}\right) \\ &\le \sigma\sqrt{L + 1}\, G\left(\frac{s - (L + 1)\mu}{\sigma\sqrt{L + 1}}\right) \end{aligned} \tag{27}$$

where the inequality follows from the fact that $G(x)$ is decreasing in x.

Once $\mathsf{E}[B]$ is derived, $\mathsf{E}[I]$ follows from (1), we have

$$\mathsf{E}[I] = \mathsf{E}[B] + \mathsf{E}[X] - \mu(L + 1)$$

For instance, from (27), taking into account $G(k) = H(k) - k$, and approximating X with a uniform distribution, we have

$$\mathsf{E}[I] = \sigma\sqrt{L + 1}\, H\left(\frac{s - (L + 1)\mu}{\sigma\sqrt{L + 1}}\right) + \frac{S - s}{2} \tag{28}$$

Alternatively, $\mathsf{E}[X]$ can follow the approximation in (26).

6. DYNAMIC (*S*, *s*) POLICY

Observe that when following any policy other than lot-for-lot, the equation in (15) governing the requirements should be modified as follows:

$$A_t = [\mu(t - L, t) + k_t\sigma(t - L, t) - Y_{t-L} - Q(t - L, t - 1)]^+ \tag{29}$$

where

$$Q(t - L, t - 1) = Q_{t-L} + \ldots + Q_{t-1}$$

replaces $A(t - L, t - 1)$ in (15). Hence, when $A_t > 0$, we have

$$\begin{aligned} A_t &= \mu(t - L, t) + k_t\sigma(t - L, t) - Y_{t-L} - Q(t - L, t - 1) \\ &= s_{t-L} - Y_{t-L} - Q(t - L, t - 1). \end{aligned}$$

That is, s_{t-L} is the reorder point, since following the DRP logic, an order ($\tilde{Q}_{t-L}$ is placed (at $t - L$) if and only if $A_t > 0$. From the above expression, A_t is the required quantity to bring the inventory position at $t - L$ back to s_{t-L}. (This is consistent with the base-stock mechanism when there are no order-size restrictions.)

Now, suppose the replenishment policy is a dynamic (*S*; *s*) rule. Specifically, when $A_t > 0$, we want to bring the inventory position to S_{t-L} ($> s_{t-L}$). Hence, an additional amount, $S_{t-L} - s_{t-L}$, is needed, and the order quantity is:

$$Q_t = A_t + S_{t-L} - s_{t-L} = S_{t-L} - Y_{t-L} - Q(t - L, t - 1) \tag{30}$$

when $A_t > 0$. (As before, $Q_t = 0$ when $A_t = 0$).

On the other hand, given any replenishment policy with a prespecified Q_t (as a function of $A_t >$ 0), we can implement this policy by setting [cf. (30)]:

$$S_{t-L} = Q_t + s_{t-L} - A_t = Q_t + Q(t - L, t - 1) + Y_{t-L}$$

when $A_t > 0$. When $A_t = 0$, we set $Q_t = 0$, which results in $S_{t-L} = s_{t-L}$, from the first equation above.

We now turn to the performance evaluation under the dynamic (*S*, *s*) rule. Note that the inventory position at $t - L$ is:

$$X_{t-L} = Y_{t-L} + Q(t - L, t) \tag{31}$$

that is, the net inventory plus the on-order quantities (orders that have been placed but have yet to arrive). In particular, when $Q_t > 0$, from (30) and (31), we have

$$S_{t-L} = Y_{t-L} + Q(t - L, t) = X_{t-L}$$

that is, after the order is placed, the inventory position is brought up to S_{t-L}.

Just as in (21), we can iterate on (13) and (14) to obtain

$$Y_t = Y_{t-L} + Q(t - L, t - 1) - D(t - L, t - 1)$$

which, upon substitution back into (14), yields:

$$I_t = [Y_{t-L} + Q(t - L, t) - D(t - L, t)]^+ = [X_{t-L} - D(t - L, t)]^+$$

and

$$B_t = [D(t - L, t) - Y_{t-L} - Q(t - L, t)]^+ = [D(t - L, t) - X_{t-L}]^+$$

These are exactly the same formulas as in (22) and (23).

Therefore, we can adapt the approximations in the stationary (*S*, *s*) model. For instance, based on the back order approximation in (27), and taking into account the formulas in (16) and (18) for the lot-for-lot case, we have the following approximation:

$$\mathsf{E}[B_t] = \sigma(t - L, t) \cdot G(k_t), \quad \text{or} \quad \mathsf{E}[B_t] = \sigma(t - L, t) \cdot G(k_t') \tag{32}$$

according to whether $A_t > 0$ or $A_t = 0$, where

$$k_t' = \frac{Y_{t-L} + Q(t - L, t - 1) - \mu(t - L, t)}{\sigma(t - L, t)}$$

which, again, can be verified as dominating k_t.

To approximate $\mathsf{E}[I_t]$, just as in the stationary case, we make use of the identity [from (22)]:

$$I_t - B_t = X_{t-L} - D(t - L, t)$$

Hence,

$$\mathsf{E}[I_t] = \mathsf{E}[B_t] + X_{t-L} - \mu(t - L, t)$$

where $\mathsf{E}[B_t]$ follows the approximation in (32), and X_{t-L} follows the expression in (31). According to the two cases in (32), and making use of the expression in (31) and the relation $H(k) = k + G(k)$, we have:

$$\mathsf{E}[I_t] = \sigma(t - L, t) \cdot H(k_t) + S_{t-L} - s_{t-L} \tag{33}$$

when $A_t > 0$ (note in this case $X_{t-L} = S_{t-L}$); and when $A_t = 0$:

$$\begin{aligned}\mathsf{E}[I_t] &= \sigma(t - L, t) \cdot H(k_t') - k_t') - k_t' \sigma(t - L, t) + X_{t-L} - \mu(t - L, t) \\ &= \sigma(t - L, t) \cdot H(k_t') \end{aligned} \tag{34}$$

Approximating $H(k_t')$ by k_t' in the above equation, we have

$$\mathsf{E}[I_t] = k_t' \sigma(t - L, t) = X_{t-L} - \mu(t - L, t)$$

The above has the intuitive interpretation that if no order is placed at $t - L$ (i.e., $R_{t_L} = Q_t = A_t = 0$), then the expected on-hand inventory at t is simply the inventory position—of which all the on-order quantities will have arrived by t—minus the demand over the lead time, period $t - L$ through period t.

To implement the above approximations requires the derivation of Y_t, which is involved in both X_t and A_t. Recursively, Y_t can be approximated by its mean:

$$\mathsf{E}[Y_t] = \mathsf{E}[I_{t-1}] - \mathsf{E}[B_{t-1}]$$

Alternatively, a cruder approximation is to forgo the distinction between the two cases $A_t > 0$ and $A_t = 0$. (Observe that this distinction is not present in the stationary case; it is averaged out in each regenerative cycle.) Specifically, ignore the k_t' case in (32), and approximate X_{t-L} by $(S_{t-L} + s_{t-L})/2$. This way, we have

$$\mathsf{E}[B_t] = \sigma(t - L, t) \cdot G(k_t) \tag{35}$$

and

$$\begin{aligned}\mathsf{E}[I_t] &= \sigma(t - L, t) \cdot H(k_t) - k_t \sigma(t - L, t) + (S_{t-L} + s_{t-L})/2 - \mu(t - L, t) \\ &= \sigma(t - L, t) \cdot H(k_t) + (S_{t-L} - s_{t-L})/2 \end{aligned} \tag{36}$$

Both are consistent with the stationary approximations in (27) and (28).

To summarize, with order-quantity restrictions, the implementation of DRP can be unified into a dynamic (S, s) control rule. Under this rule, the constrained order quantity needed at the beginning of period t (Q_t), and hence the recommended order quantity at the beginning of period $t - L$ $(\tilde{Q}_{t-L})$, should be

$$\tilde{Q}_{t-L} = Q_t = [A_t + S_{t-L} - s_{t-L}] \cdot 1[A_t > 0]$$

where 1[·] denotes the indicator function, and

$$A_t = [s_{t-L} - Y_{t-L} - Q(t - L, t - 1)]^+$$

following (29), with

$$s_{t-L} = \mu(t - L, t) + k_t\sigma(t - L, t)$$

6.1. Setting Safety Stock Levels

In the preceding discussions, the safety stock level, Σ_t, is assumed as given. Here we discuss one approach to set the safety stock levels, which is of particular importance because of its wide usage. It sets safety stock levels based on achieving a target fill rate.

For ease of discussion, here we assume stationary demand and omit the time index t wherever possible. Suppose the service requirement is that the fraction of demand back ordered should be limited to $1 - \beta$, where β is the required fill rate.

To characterize the fraction of back ordered demand, we need to pick a "typical" time frame. In the standard inventory literature, this is taken to be the time between two consecutive orders, known as a regenerative cycle (when demands are independent and identically distributed). The required fraction is then the ratio of the average number of back orders to the average number of demand units, both over the regenerative cycle.

In the stationary (S, s) model, the average number of back orders, following (27), is approximated by

$$\sigma\sqrt{L + 1}G\left(\frac{s - (L + 1)\mu}{\sigma(L + 1}\right)$$

whereas the average demand per cycle is

$$\left(\frac{S - s}{\mu} + \frac{m_2}{2\mu^2}\right)\mathsf{E}(D) = S - s + \frac{m_2}{2\mu}$$

where the first factor on the left hand side is the expected cycle length $\mathsf{E}[T_s] = 1 + M(S - s)$ (refer to Section 5.1) with $M(S - s)$ approximated by the linear asymptote in (25). Note that the right-hand side above is nothing but the expected order quantity; in particular, $m_2/2\mu$ is the expected undershoot.

Therefore, based on the above, under the dynamic (S, s) rule, k_t should be the solution to:

$$\sigma(t - L, t)G(k_t) = (1 - \beta)\left[\Delta_{t-L} + \frac{\mathsf{E}(D^2_{t-L})}{2\mathsf{E}(D_{t-L})}\right] \tag{37}$$

where $\Delta_{t-L} = S_{t-L} - s_{t-L}$ is assumed given.

The lot-for-lot rule is equivalent to $S_{t-L} = s_{t-L}$, or $\Delta_{t-L} = 0$. Also, since in every period an order is placed, the order quantity is simply equal to the demand. Hence, the equation in (37) is reduced to:

$$\sigma(t - L, t)\ G(k_t) = (1 - \beta)\mathsf{E}(D_{t-L}) \tag{38}$$

Note that the equations in (37) and (38) are easily solved through Newton's method (e.g., Press et al. 1994). Write the equations in the form of $G(k) = c$. The Newton's iteration, indexed by the superscript (n), is as follows:

$$k^{(n+1)} = k^{(n)} - \frac{G(k^{(n)}) - c}{G'(k^{(n)})}$$

When demand follows a normal distribution, $G'(k) = -1 + \Phi(k)$, where Φ denotes the distribution function of the standard normal variate.

7. MULTI-STAGE MODELS

The results presented so far for the single-stage model extend to a more general distribution network. For ease of exposition, we consider a model that consists of a central stocking facility (e.g., a warehouse or depot) supplying a set of local stocking facilities (e.g., retailers or outlets), each of which, in turn, supplies its own customer demand. As in the single-stage models, our focus here is on performance evaluation via simple approximations. We do not address the issue of cost optimization, for which many models can be found in the survey article by Federgruen (1993).

Suppose the central warehouse is numbered as stage 0 and the set of retailers, numbered as stages $1, \ldots, c$. Quantities that relate to these stages will be subscripted or superscripted (if there is already a subscript for time) by their stage indices. For instance, the demand stream to retailer i, $i = 1, \ldots, c$, is $\{D_t^i, t = 1, 2, \ldots\}$. Assume the demands are independent among the retailers, and independent over the time periods.

Suppose the lead time at each retailer i is L_i periods, a deterministic constant, for $i = 1, \ldots, c$. For instance, this can be the transportation time for a replenishment order to travel from the warehouse to the retailer; or, in the case where stage i is a plant, the cycle time to build a customer order. Suppose the lead time at the warehouse is L_0.

The analysis of each retailer i follows the same discussion as before. In particular, write

$$s_t^i = \mu_i(t, t + L_i) + k_{t+L_i}^i \sigma_i(t, t + L_i)$$

and if a set of target fill rates at each of the retailers, β_i, $i = 1, \ldots, c$, has been specified, then the safety factor $k_{t+L_i}^i$ is obtained from solving the following equation. (Refer to Eqs. (37) and (38)).

$$\sigma_i(t, t + L_i)G(k_{t+L_i}^i) = (1 - \beta_i)\left[\Delta_t^i + \frac{\mathsf{E}(D_t^i)^2)}{2\mathsf{E}(D_t^i)}\right]$$

(where $\Delta_t^i := S_t^i$ is assumed given); or, from

$$\sigma_i(t, t + L_i)G(k_{t+L_i}^i) = (1 - \beta_i)\mathsf{E}(D_t^i)$$

in the special case of i following a lot-for-lot rule.

The expected number of back orders is:

$$\mathsf{E}[B_t^i] = \sigma_i(t - L_i, t)G\left(\frac{s_{t-L}^i - \mu_i(t - L_i, t)}{\sigma_i(t - L_i, t)}\right)$$

corresponding to the k_t case in (32), that is, when $A_t^i > 0$. For the other case ($A_t^i = 0$), the expression is similar, corresponding to the k_t' case in (32). The expected on-hand inventory is where $X_{t-L_i}^i$ denotes the inventory position.

To analyze the warehouse, we follow the standard notion of echelon stock. That is, we aggregate stage 0, along with stages $i = 1, \ldots, c$, into a single system, indexed by E (for echelon). This aggregated system supplies the superposition of all c demand streams. Note that each demand process D_t^i still has its own lead time L_i. Let

$$L_{\max} = \max\{L_i, i = 1, \ldots, c\},$$

Then (22) and (23) should be modified, with L replaced by $L_E = L_0 + L_{\max}$.

Therefore, the lower threshold value for this aggregated system should be set, in period t, as follows:

$$s_t^E = \mu_E(t, t + L_E) + k_{t+L_E}^E \sigma_E(t, t + L_E)$$

where

$$\mu_E(t, t + L_E) = \sum_{i=1}^{c} \mu_i(t, t + L_E)$$

$$\sigma_E(t, t + L_E) = \left[\sum_{i=1}^{c} \sigma_i^2(t, t + L_E)\right]^{1/2}$$

As for the echelon safety factor, $k_{t+L_E}^E$, it can either be prespecified by the user or, if the safety stock levels are to be set automatically based on target fill rates specified at the retail level, it can be obtained from the solution to:

$$\sigma_E(t, t + L_E)G(k_{t+L_E}^E) = (1 - \beta_{\max})(\Delta_t^E + \gamma_t^E)$$

with $\Delta_t^E = S_t^E - s_t^E$ assumed given, $\beta_{\max} = \max\{\beta_i, i = 1, \ldots, c\}$, where β_i is the target fill rate at retailer i, and

$$\gamma_t^E = \frac{\mathsf{E}\left[\left(\sum_{i=1}^{c} D_t^i\right)^2\right]}{2\mathsf{E}\left(\sum_{i=1}^{c} D_t^i\right)} = \frac{\sum_{i=1}^{c} (\sigma_t^i)^2 + \left(\sum_{i=1}^{c} \mu_t^i\right)^2}{2\sum_{i=1}^{c} \mu_t^i}$$

That is, stage 0 should place an order at the beginning of period t if and only if the *echelon* inventory position falls below s_t^E. This echelon inventory position is the inventory position with the aggregated system considered as a single stage. It includes all the on-hand inventory at both the warehouse and all the retailers, plus any replenishment orders placed by the warehouse that have not yet arrived, minus any customer demand that is back ordered at all the retailers. Hence, this echelon inventory position will only change whenever there is a customer order arriving at a retailer and whenever the warehouse places a replenishment order. It will not change when a retailer takes supply from or registers a back order with the warehouse. When stage 0 issues a replenishment order, say at the beginning of period t, the order quantity is to bring the echelon stock position to $S_t^E = \Delta_t^E + s_t^E$. Hence, $\Delta_t^E + \gamma_t^E$ is the estimate of this order quantity.

Estimates of $\mathsf{E}[B_t^E]$ and $\mathsf{E}[I_t^E]$, like those at the retailers, follow in the same vein as those in Section 6 (refer to, in particular, the summary at the end of that section).

The relationship between Q_t and S_{t-L} follows (30); in particular, when $A_t > 0$, $Q_t = S_{t-L} - Y_{t-L} - Q(t - L, t - 1)$; and when $A_t = 0$, $Q_t = 0$ and $S_{t-L} = s_{t-L}$. The estimates on the expected on-hand inventory and back orders (for period t) follow (33), (34) and (32), respectively; or alternatively, follow (36) and (35), respectively.

8. ASSEMBLE-TO-ORDER SYSTEMS

An assemble-to-order (ATO) system is a hybrid model of build-to-stock at the component (subassembly) level and assemble-to-order for the end product. In an ATO system, typically, the components take a substantial lead time to build, whereas the time it takes to assemble all the components into the final product is often negligible. Hence, keeping stock at the component level improves response-time performance, whereas not keeping any end-product inventory reduces inventory cost and maximizes the flexibility for customization. A good example of an ATO system is the production of a PC (personal computer). Other examples include fast-food operations and many mail-order or e-commerce services.

We consider an ATO system of m different items (components) and a single end product. Let $\mathcal{I} = \{1, 2, \dots, m\}$ denote the set of all items. Without loss of generality, assume each end product consists of exactly one unit of each item in $\mathcal{I}$. (If the end product requires multiple units from some item, simply redefine the unit for that component and adjust inventory and lead time accordingly.) Customer demand for the end product follows a stationary Poisson process, denoted $\{A(t), t \geq 0\}$, with rate λ.

Demands are filled on a first-come-first-served (FCFS) basis. If there is positive on-hand inventory for all components upon a demand arrival, the demand is filled immediately (since the time to assemble the components into the end-product is negligible). On the other hand, if there is a stockout at one or more of the component inventory, then the demand is backlogged until the stockout inventory is replenished.

For each unit of item i, we assume the production lead times are i.i.d. random variables with a common distribution function G_i, and with L_i denoting the associated random variable, and $\mathsf{E}[L_i] = \ell_i$. Assume the lead times are independent among the items.

The inventory of each item i is controlled by a base-stock policy, with R_i being the base-stock level. Since the end product consists of a single unit of each component, the base-stock policy implies that every demand will trigger *simultaneously* the production/replenishment of one unit of each component. Therefore, as in the model of Section 2.1, for each item i, the number of units in process at any time t, denoted $X_i(t)$, is equal to the number of jobs in service in an $M/G_i/\infty$ queue, for $i = 1, \dots, m$. Note, however, that these m queues are driven by a common Poisson arrival process $\{A(t)\}$, and hence are *not* independent.

For any time t, the performance measures of interest are the on-hand inventory and the number of back orders:

$$I_i(t) = [R_i - X_i(t)]^+, \quad \text{and} \quad B_i(t) = [X_i(t) - R_i]^+ \tag{39}$$

Note that from the FCFS rule, the number of back ordered demand (for the end product) is:

$$B(t) = \max_{1\leq i\leq m} B_i(t) = \max_{1\leq i\leq m} [X_i(t) - s_i]^+ \tag{40}$$

Let X_i, I_i, B_i, and B denote the corresponding steady-state limits of the above quantities.

Let f_i and f denote the fill rates for item i and for the end product, respectively. Then, due to the property that Poisson arrivals see time average (PASTA) (e.g., Wolff 1989), we have

$$f_i = \mathsf{P}(I_i > 0), \quad \text{and} \quad f = \mathsf{P}(I_1 > 0, \ldots, I_m > 0)$$

Note that the item-based performance measures f_i, I_i, and B_i depend only on the marginal distribution of X_i for each i, while the product-based performance measures f and B depend on the joint distribution of X_i, $i = 1, \ldots, m$.

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

Let $N(a)$ denote a Poisson random variable with parameter (mean) a, along with the following notation:

$$p(n|a) = \mathsf{P}[N(a) = n] = \frac{a^n}{n!}e^{-a}$$

$$P(n|a) = \mathsf{P}[N(a) \leq n] = \sum_{k=0}^{n} p(k|a)$$

$$\overline{P}(n|a) = \mathsf{P}[N(a) > n] = \sum_{k=n+1}^{\infty} p(k|a) = 1 - P(n|a)$$

for $n = 0, 1, \ldots$ Given $R \geq 1$, an integer, we have

$$b(R|a) = \mathsf{E}[N(a) - R]^+ = \sum_{n=1}^{\infty} np(n + R|a) = \sum_{n=R}^{\infty} \overline{P}(n|a) = a - \sum_{n=0}^{R-1} \overline{P}(n|a)$$

The third equality above follows from rearranging the terms, and the fourth equality applies the identity $\Sigma_{n=0}^{\infty} \overline{P}(n|a) = a$. Note that $b(0|a) = \mathsf{E}[N(a)] = a$ and that $b(R|a)$ is decreasing in R, with $b(+\infty|a) = 0$.

Since X_i is equal in distribution to $N(\lambda\ell_i)$, we can express the item-based performance measures as follows:

$$f_i = \mathsf{P}[X_i \leq R_i - 1] = \mathsf{P}(R_i - 1|\lambda\ell_i)$$

$$\mathsf{E}[B_i] = b(R_i|\lambda\ell_i)$$

$$\mathsf{E}[I_i] = R_i - \mathsf{E}[X_i] + \mathsf{E}[B_i] = R_i - \lambda\ell_i + \mathsf{E}[B_i]$$

The last equality above follows from $I_i - B_i = R_i - X_i$, which in turn follows from (39).

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

$$X_i(t) = N_0(t) + N_i(t)$$

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

Consider a given $t > 0$. Suppose $A(\mathrm{t}) = n$. Then, it is well known (e.g., Ross 1996) (20) that the n (unordered) arrivals follow an i.i.d. uniform distribution in $[0, t]$ and that $(N_0(t), N_1(t), N_2(t))$ follows a multinomial distribution (of n objects, into four categories), with the following probabilities:

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

$$p_1(t) = \int_0^t \overline{G}_1(t - x)G_2(t - x)(dx/t) = \int_0^t \overline{G}_1(x)G_2(x)(dx/t)$$

$$p_2(t) = \int_0^t G_1(t - x)\overline{G}_2(t - x)(dx/t) = \int_0^t G_1(x)\overline{G}_2(x)(dx/t)$$

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

$$\begin{aligned}
&\mathsf{P}[\mathrm{N}_0(t) = n_0, N_1(t) = n_1, N_2(t) = n_2] \\
&\quad = \sum_{n \geq n_0+n_1+n_2} \frac{n!}{n_0!n_1!n_2!(n - n_0 - n_1 - n_2)!} [p_0(t)]^{n_0}[p_1(t)]^{n_1}[p_2(t)]^{n_2} \\
&\qquad \cdot [1 - p_0(t) - p_1(t) - p_2(t)]^{n-n_0-n_1-n_2} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \\
&\quad = \frac{[\lambda t p_0(t)]^{n_0}[\lambda t p_1(t)]^{n_1}[\lambda t p_2(t)]^{n_2}}{n_0!n_1!n_2!} \cdot \exp[-\lambda t(p_0(t) + p_1(t) + p_2(t))] \qquad (41)
\end{aligned}$$

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

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

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

From (40), we have

$$\begin{aligned}
&\mathsf{P}[B(t) \leq x] \\
&\quad = \mathsf{P}[X_1(t) \leq x + s_1, X_2(t) \leq x + s_2] \\
&\quad = \mathsf{P}[N_0(t) + N_1(t) \leq x + s_1, N_0(t) + N_2(t) \leq x + s_2] \\
&\quad = \sum_{n_0=0}^{(s_1 \wedge s_2)+x} \mathsf{P}[N_0(t) = n_0]\mathsf{P}[N_1(t) \leq x + s_1 - n_0, N_2(t) \leq x + s_2 - n_0] \\
&\quad = \sum_{n_0=0}^{(s_1 \wedge s_2)+x} p(n_0|\lambda t p_0(t))\mathsf{P}(x + s_1 - n_0|\lambda t p_1(t))\mathsf{P}(x + s_2 - n_0|\lambda t p_2(t))
\end{aligned}$$

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

The above analysis extends readily to $m > 2$. In particular, there will be a total of $2^m - 1$ independent Poisson random variables involved: $N_{\mathbb{S}}(t)$, for $\mathbb{S} \subset \mathcal{J}$, representing the number of jobs that are still in process at time t with the queues $i \in \mathbb{S}$, but have been completed with the queues $j \in \mathcal{J}\backslash\mathbb{S}$. For each $i = 1, \ldots, m$, we can write

$$X_i(t) = \sum_{\mathbb{S}:i\in\mathbb{S}} N_{\mathbb{S}}(t)$$

That is, $X_i(t)$ can be expressed as the sum of $2^m - 1$ independent Poisson random variables. Hence, in principle, all the performance measures at time t can be exactly evaluated based on the distributions of the independent Poisson random variables via $X_i(t)$'s. The exponential growth (w.r.t. m), however, makes this impractical, even for systems with a moderately large number of components.

An exception is the special case of deterministic lead times, that is, $L_1 \equiv \ell_1$, $L_2 \equiv \ell_2$. Without loss of generality, assume $\ell_1 < \ell_2$. For any fixed t, there are three cases.

Case 1: $t > \ell_2$. In this case,

$$
\begin{aligned}
p_0(t) &= \int_0^t 1[\ell_1 \geq s]1[\ell_2 \geq s](ds/t) \\
&= \int_0^t 1[\ell_1 \geq s](ds/t) = \ell_1/t
\end{aligned}
$$

Similarly,

$$
p_1(t) = \int_0^t 1[\ell_1 \geq s]1[\ell_2 \leq s](ds/t) = 0
$$

and

$$
\begin{aligned}
p_2(t) &= \int_0^t 1[\ell_1 \leq s]1[\ell_2 \geq s](ds/t) \\
&= \int_0^t 1[\ell_1 \leq s \leq \ell_2](ds/t) = (\ell_2 - L_1)/t
\end{aligned}
$$

Since $p_1(t) = 0$, $N_1(t) \equiv 0$, we have

$$
X_1(t) = N_0(t) \tag{42}
$$

$$
X_2(t) = N_2(t) + N_0(t) \tag{43}
$$

Case 2: $l_1 < t \leq \ell_2$. In this case, we have $p_2(t) = (t - \ell_1)/t$, while $p_0(t) = \ell_1/t$ and $p_1(t) = 0$ stay the same as in Case 1; and (42) and (43) hold.

Case 3: $t \leq \ell_1$. In this case, we have $p_0(t) = 1$, and $p_1(t) = p_2(t) = 0$. So none of the arrived jobs is completed at either queue at time t. Therefore,

$$
X_1(t) = X_2(t) = N_0(t).
$$

Thus, in the special case of deterministic lead times, there are only $m - 1$ independent Poisson random variables involved, as opposed to $2^m - 1$ in the case of random lead times.

The steady-state performance evaluation can be conducted in a similar manner, as indicated in the next section. Denote

$$
\begin{aligned}
\theta_0 &= \lim_{t\to\infty} tp_0(t) = \int_0^\infty \overline{G}_1(x)\overline{G}_2(x)\,dx \\
\theta_1 &= \lim_{t\to\infty} tp_1(t) = \int_0^\infty \overline{G}_1(x)G_2(x)\,dx \\
\theta_2 &= \lim_{t\to\infty} tp_2(t) = \int_0^\infty G_1(x)\overline{G}_2(x)\,dx
\end{aligned}
$$

Let $t \to \infty$ (41), and write $N_i = N_i(\infty)$ for $i = 0, 1, 2$. We have

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

Thus, N_i, $i = 0, 1, 2$, are independent Poisson random variables with parameters $\lambda\theta_i$, $i = 0, 1, 2$, respectively.

Now, consider the steady-state limit of $(X_1(t), X_2(t))$, denoted (X_1, X_2). First notice the following:

$$
\theta_0 + \theta_1 = \ell_1, \quad \theta_0 + \theta_2 = \ell_2
$$

From the infinitely divisible property of the Poisson distribution, we can write

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

The three Poisson random variables, N_0, N_1N_2, involved in the above expressions are independent of one another. Also, X_1 and X_2 are correlated through a common random component $N(\lambda\theta_0)$.

9. KANBAN CONTROL

9.1. The Basic Model

Kanban control refers to a control rule that limits the inventory level, including both WIP and finished goods to K units, with K a constant parameter. Specifically, there are K cards in the system; each card is attached to a job that is in waiting or in process, or to a job that is completed (waiting to supply demand). In other words, a card is needed to admit an order into the system. When all K cards are exhausted, there are three possibilities:

1. If all K cards are attached to completed jobs, then production is suspended.
2. If all K cards are attached to jobs in waiting or in process (i.e., there is no completed job), then any further arrival (demand) will be blocked, i.e., no more jobs are admitted into the system.
3. The situation in between is that some of the K cards are attached to completed jobs while others are attached to jobs in waiting or in process. Then, any arriving demand will be supplied by one of the completed jobs, with its card detached and given to another new job (representing the outstanding order).

This clearly relates to the base-stock control mechanism. In particular, $K = R$—any unit of demand will trigger production (or replenishment), whereas when the finished goods inventory reaches K, production will be suspended. Kanban, however, has the additional feature of blocking arrivals (of demand) when the on-hand inventory drops down to zero, that is, when all K cards are associated with outstanding orders, which is the situation in (2) above. Hence, kanban corresponds to a *finite* queueing system with K being the buffer capacity—the upper limit on the total number of jobs allowed in the system.

With this in mind, in particular, with $R = K$, the models in Section 2.1 should be modified to $M/G/\infty/K$ and $M/M/1/K$ queues (e.g., Wolff 1989). In terms of the distributions of N, I, and B, this amounts to a renormalization: dividing those probabilities in the last section by $\mathsf{P}[N \geq K]$. The inventory and back order expressions in (1) remain valid, with $R = K$. For instance, in the $M/G/\infty/K$ case, we have

$$\mathsf{P}[I = 0] = \mathsf{P}[N = K] = \frac{\rho^K}{K!} e^{-\rho} \left[\sum_{n=0}^{K} \frac{\rho^n}{n!} e^{-\rho} \right]^{-1}$$
$$\mathsf{P}[I = n] = \frac{\rho^{R-n}}{(R-n)!} e^{-\rho} \left[\sum_{n=0}^{K} \frac{\rho^n}{n!} e^{-\rho} \right]^{-1}, \ n = 1, \ldots, R$$

and

$$\mathsf{P}[B = 0] = 1,$$

since $N \leq K$.

9.2. Generalized Kanban Control

Note that in the above kanban control mechanism, the cards control both admission (of demand arrivals) into the system and the upper limit of finished goods inventory. A more general kanban control mechanism is one that uses two types of cards. In addition to the usual type, which now (only) controls admission, there is a second type of cards controlling finished goods inventory. Let K and R, with $K \geq R$, denote, respectively, the number of cards for these two types, referred to, respectively, as kanbans and production cards. Specifically, every job within the system has a kanban attached to it. In addition, to be processed by the server, a job also requires a production card. Both the kanban and the production card stay with the completed job until it supplies a demand. At that point, the released kanban admits a new order into the system, while the released production card authorizes the service (production) of another waiting order. (Note that the above kanban control with two types of cards can be shown to be equivalent to the kanban control with three types of cards, the so-called (a, b, k) model in Glasserman and Yao 1994b, 1996.) This way, the total number of jobs in the system is limited to K, while the total number of *completed* jobs (that are waiting for demand) is limited to R. Since now $R \leq K$, back orders are allowed, up to a limit of $K - R$.

Therefore, what is needed for this more general kanban control is the $M/G/\infty/K$ and $M/M/1/K$ models, in connection with the relations in (1)—with $R \leq K$. The earlier distributions for I and B can be modified accordingly. For instance, in the $M/M/1/K$ case, we have

$$\mathsf{P}[I = 0] = \mathsf{P}[N \geq R] = \frac{\rho^R - \rho^{K+1}}{1 - \rho^{K+1}}$$

$$\mathsf{P}[I = n] = \mathsf{P}[N = R - n] = \frac{\rho^{R-n}(1 - \rho)}{1 - \rho^{K+1}},\ n = 1, \ldots, R$$

and

$$\mathsf{P}[B = 0] = \mathsf{P}[N \leq R] = \frac{1 - \rho^{R+1}}{1 - \rho^{K+1}}$$

$$\mathsf{P}[B = n] = \mathsf{P}[N = R + n] = \frac{\rho^{R+n}(1 - \rho)}{1 - \rho^{K+1}},\ n = 1, \ldots, K - R$$

The associated expectations can be derived as follows:

$$\mathsf{E}[I] = \frac{R - (R + 1)\rho + \rho^{R+1}}{(1 - \rho)(1 - \rho^{K+1})}$$

$$\mathsf{E}[B] = \frac{\rho^{R+1} - (K - R + 1)\rho^{K+1} + (K - R)\rho^{K+2}}{(1 - \rho)(1 - \rho^{K+1})}$$

10. NETWORK OF INVENTORY QUEUES

Now consider an inventory/distribution network, also known as a supply chain. Each node in the network represents a stocking location. Suppose a base-stock control policy is followed at each node. With the discussion above, we can adapt the standard decomposition approach in analyzing queueing networks to study this inventory network.

To be specific, focus on a particular node, j, in the network. Suppose node j has a single upstream node i, which supplies any replenishment orders from j. Since base-stock control is followed throughout the network, each node is driven directly by the external demand process, which we continue to assume to be a Poisson process with rate λ.

Suppose j is modeled as an $M/G/\infty$ queue. Then, the only difference from the analysis in Section 2 is that the service time (lead time), L_j, needs to be prolonged whenever there is a stockout at node i. Hence, the modified service time is:

$$\tilde{L}_j = L_j + \tau_i, \text{ w.p. } \Phi(k_i) \tag{44}$$

While $\tilde{L}_j = L_j$ w.p. $\Phi(i_i)$. Here, τ_i is the extra delay at node i, until the next job (outstanding order) is completed. Following the analysis in Ettl et al. (2000), we can approximate τ_i as follows (provided node i is also modeled as an $M/G/\infty$ queue):

$$\tau_i = \frac{L_i\mathsf{E}(B_i)}{R_i\overline{\Phi}(k_i)} \tag{45}$$

Next, suppose both i and j are modeled as single-server queues, say $M/M/1$. Then, the only occasion in which the service time at i needs to be prolonged is when j has a stockout and server i is forced to become idle. Obviously, the probability is dominated by the stockout probability at node j, which is equal to $\rho_j^{R_j}$. In the single-server queue model, stability requires $\rho_j < 1$, hence this probability is quite negligible when R_j is reasonably large and ρ_j is not too close to 1. In this case, we can simply forgo the adjustment. Otherwise, an adjustment similar to the one in (44) is to add $\tau_i = L_i$ to L_j w.p. $\rho_j^{R_j}$. (However, this might result in overcompensation; in particular, the associated probability, $\rho_j^{R_j}$, could be too high.)

Suppose node j interacts directly with external customer demand, and follows a generalized kanban control as outlined in Section 9.2. Then a proportion of the demand will be blocked and lost when all K_j cards are occupied. This is equal to the probability $\mathsf{P}[N_j = K_j]$, where N_j follows the

$M/G/\infty/K_j$ or the $M/G/1/K_j$ model, with the adjusted service time $\tilde{L}_j$ as discussed above. Note that in this case we have a combination of back order and lost sales for external demand.

The simple Poisson demand arrival process assumed here can be readily extended to more involved processes, such as those that involve batches, like the model in Ettl et al. (2000), which makes use of the queueing results in Liu et al. (1990). The same applies to more general lead time distributions and multiple servers; approximate queueing models can be adapted; refer to Buzacott and Shanthikumar (1993).

Optimization models can be formulated based on these approximate inventory-queue models, Ettl et al. (2000) being one such example. As the basic relations in (1) serve as building blocks of the network, one should be able to establish structural properties of the objective function, based on properties of the functions in (1), which are convex in R and submodular in (R, N). These are useful properties in identifying efficient solution algorithms.

In this type of decomposition, each decomposed node is driven by exactly the same external demand processes, due to the base-stock control mechanism. This is quite different from the usual queueing network decomposition, where it is standard to assume that the decomposed queues are independent of each other, which is indeed the case in the special case of product-form networks. In the network of inventory queues, the dependence among the queues should be maximal since the queues are all driven by the same arrival processes. In this regard, the network behaves more like the ATO system in Section 8, where each unit of demand consists of a set of components; and hence, a demand arrival will simultaneously trigger the production at all component queues.

Therefore, to conclude this section, let us examine more closely the issue of stockout in the ATO system and its relation with customer order service. Consider multiple demand streams, indexed by $m \in \mathcal{M}$. Let $\mathcal{S}_m$ denote the set of components required to assemble one unit of end product for type m demand. Let α be the required service level, defined here as the off-shelf availability of all the components required to assemble a unit of type m product, for any m. Let E_i denote the event that component i is out of stock. Then we require, for each end product $m \in \mathcal{M}$,

$$\mathsf{P}[\cup_{i\in\mathcal{S}_m} E_i] \leq 1 - \alpha$$

From the well-known inclusion–exclusion formula (e.g., Ross 1996):

$$\mathsf{P}[\cup_{i\in\mathcal{S}_m} E_i] = \sum_i \mathsf{P}(E_i) - \sum_{i<j} \mathsf{P}(E_i \cap E_j) + \sum_{i<j<k} \mathsf{P}(E_i \cap E_j \cap E_k) - \ldots$$

we have, as an approximation,

$$\mathsf{P}[\cup_{i\in\mathcal{S}_m} E_i] \cong \sum_{i\in\mathcal{S}_m} \mathsf{P}(E_i) = \sum_{i\in\mathcal{S}_m} \bar{\Phi}(k_i) \leq 1 - \alpha \tag{46}$$

Note that $\bar{\Phi}(k_i)$ is the stockout probability of component i, using a normal approximation [c.f. (4)].

There is another way to arrive at the above inequality. Suppose we express the service requirement as follows:

$$\prod_{i\in\mathcal{S}_m} \Phi(k_i) \geq \alpha, \quad m \in \mathcal{M} \tag{47}$$

Note that the left-hand side in (47) is, in fact, a *lower bound* of the no-stockout probability of the set of components in $\mathcal{S}_m$ that is required to assemble the end product m, that is, it is a lower bound of the desired off-shelf availability. This claim (of a lower bound) can be argued by using stochastic comparison techniques involving the notion of *association.* (Refer to, e.g. Ross 1996 for background materials.) Intuitively, since the component inventories are driven by a common demand stream $\{D_m(t)\}$, and hence positively correlated, the chance of missing one or several components must be less than when the component inventories are independent, which is what is assumed by the product on the left-hand side of (47).

Since

$$\prod_{i\in\mathcal{S}_m} \Phi(k_i) = \prod_{i\in\mathcal{S}_m} [1 - \bar{\Phi}(k_i)] \cong 1 - \sum_{i\in\mathcal{S}_m} \bar{\Phi}(k_i) \tag{48}$$

where the approximation works best when the stockout probability $\bar{\Phi}(k_i)$ is small, for all components $i \in \mathcal{S}_m$. [Note the approximation in (48) is analogous to the one in (46).] Combining (48) with (47), we arrive at the same inequality in (46).

We can now relate the above off-shelf availability requirement to the standard customer service requirements expressed in terms of response times, W_m, the time it takes to fill a customer order (of type $m \in \mathcal{M}$). Suppose the required service level of type m demand is

$$\mathsf{P}[W_m \leq w_m] \geq \alpha, \quad m \in \mathcal{M} \tag{49}$$

where w_m's are given data. This is the type of service requirement considered in Ettl et al. (2000).

Consider type m demand. We have the following two cases:

1. When there is no stockout at any store $i \in \mathcal{S}_m$—denoting the associated probability as $\pi_{0m}(t)$, the delay is simply L_m^{out}, the outbound lead time—time to process the order, assemble the product, and deliver it to the customer.
2. Suppose there is a stockout at a store $i \in \mathcal{S}_m$. Denote the associated probability as $\pi_{im}(t)$. Then the delay becomes $L_m^{\text{out}} + \tau_i$, where τ_i is the additional delay before the stocked-out component becomes available [cf. (45)].

Hence, we can write

$$\begin{aligned}\mathsf{P}[W_m \leq w_m] &\cong \pi_{0m}(t)\mathsf{P}[L_m^{\text{out}} \leq w_m] + \sum_{i \in \mathcal{S}_m} \pi_{im}(t)\mathsf{P}[L_m^{\text{out}} + \tau_i \leq w_m] \\ &= \left[\prod_{i \in \mathcal{S}_m} \Phi(k_i)]\mathsf{P}[L_m^{\text{out}} \leq w_m\right] + \sum_{i \in \mathcal{S}_m} \overline{\Phi}(k_i)\mathsf{P}[L_m^{\text{out}} + \tau_i \leq w_m]\end{aligned} \tag{50}$$

Note that in the approximation above, we have ignored the probability of two or more components stocking out at the same time [in the same spirit as in (46) and (48)].

In most applications, it is reasonable to expect $L_m^{\text{out}} \leq w_m$. For instance, this is the case when the outbound lead time L_m^{out} is nearly deterministic, and the delay limit w_m is set to be safely larger than L_m^{out}. Furthermore, τ_i can be estimated based on (45). Therefore, if we set the delay limit w_m such that $w_m \geq L_m^{\text{out}} + \tau_i$ with probability one, for all $i \in \mathcal{S}_m$, then the response-time serviceability in (50) can be met at nearly 100% [taking into account (48)]. In other words, aiming for a high off-shelf availability (say, 90–95%) will usually enable us to set a reasonable response-time limit (w_m), and to achieve a near-100% response-time service level.

11. BIBLIOGRAPHICAL NOTES

Buzacott and Shanthikumar (1993) is a rich source of exact and approximate queueing models of production-inventory systems, including generalizations of the models in Section 2.1 and Section 9. Kanban control is studied in detail from a discrete-event systems point of view in Glasserman and Yao (1994a, b, 1996); in particular, the dynamics are modeled as generalized semi-Markov processes, and structural properties such as monotonicity, concavity, and line reversibility (symmetry) are exploited to solve the optimal allocation of kanbans among the production stages. The PAC (production authorization cards) scheme in Buzacott and Shanthikumar (1993) provides a unified modeling framework for many production control schemes, including kanban and MRP.

The materials in Sections 3 through 7 are drawn from Feigin et al. (2000), which also contains a critique of DRP, including numerical examples.

The discussion on ATO systems in Sections 8 and 10 draws materials from Cheng et al. (2000) and Song and Yao (2000). Both papers also have extensive treatments of optimization models, with the latter focusing on queueing analysis and stochastic bounds (also refer to Song 1998; Song et al. 1999), whereas the former focuses on normal approximations and industrial (PC manufacturing) applications. Glasserman and Wang (1998) study ATO systems using a large deviations-based approach, deriving asymptotic results; also refer to Glasserman (1997).

The decomposition-based approach overviewed in Section 10 was first developed in Ettl et al. (2000) for the purpose of performance evaluation and optimization of a large-scale enterprise supply chain. (Refer to Lee and Billington 1986 for an earlier, related work on modeling supply chains.) A related network model, for semiconductor fabrication, appeared in Connors et al. (1996). Also refer to Buzacott and Shanthikumar (1993) for other network models using decomposition-based approximations.

An important topic that we have not discussed in this chapter is the incorporation of quality control into the modeling of production-inventory system (refer to Chen et al. 2000; Yao and Zheng 1999a, b), where the emphasis is on coordinating production and quality control (e.g., inspection), with quality adding a new dimension to the usual inventory-service trade-off.

REFERENCES

Buzacott, J. A., and Shanthikumar, J. G. (1993), *Stochastic Models of Manufacturing Systems,* Prentice Hall, Englewood Cliffs, NJ.

Chen, J., Yao, D. D., and Zheng, S. (2000), "Optimal Replenishment and Rework with Multiple Unreliable Supply Sources," *Operations Research* (forthcoming).

Cheng, F., Ettl, M., Lin, G. Y., and Yao, D. D. (2000), "Inventory-Service Optimization Configure-to-Order Systems: From Machine Type Models to Building Blocks," IBM Research Report.

Clark, A. J., and Scarf, H. (1960), "Optimal Policies for a Multi-Echelon Inventory Problem," *Management Science*, Vol. 6, pp. 475–490.

Connors, D., Feigin, G., and Yao, D. D. (1996), "A Queueing Network Model for Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, pp. 412–427.

Ettl, M., Feigin, G., Lin, G. Y., and Yao, D. D. (2000), "A Supply Network Model with Base-Stock Control and Service Requirements," *Operations Research* (forthcoming).

Federgruen, A. (1993), "Centralized Planning Models for Multi-Echelon Inventory Systems under Uncertainty," in *Logistics of Production and Inventory,* S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin, Eds., North-Holland, Amsterdam, pp. 133–173.

Feigin, G. E., Katircioglu, K., and Yao, D. D. (2000), "Distribution Resource PlanningSystems: A Critique and Enhancement," IBM Research Report.

Glasserman, P. and Yao, D. D. (1994a), *Monotone Structure in Discrete-Event Systems*, John Wiley & Sons, New York.

Glasserman, P., and Yao, D. D. (1994b), "A GSMP Framework for the Analysis of Production Lines," in *Stochastic Modeling and Analysis of Manufacturing Systems,* D. D. Yao, Ed., Springer, New York.

Glasserman, P., and Yao, D. D. (1996), "Structured Buffer Allocation Problems," *Discrete Event Dynamic Systems: Theory and Applications*, Vol. 6, pp. 9–42.

Glasserman, P. (1997), "Bounds and Asymptotics for Planning Critical Safety Stocks," *Operations Research*, Vol. 45, pp. 244–257.

Glasserman, P. and Wang, Y. (1998), "Leadtime–Inventory Tradeoffs in Assemble-to-Order Systems," *Operations Research*, Vol. 46, pp. 858–871.

Li, L. (1992), "The Role of Inventory in Delivery-Time Competition," *Management Science*, Vol. 38, pp. 182–197.

Lee, H., and Billington, C. (1986), "Material Management in Decentralized Supply Chains," *Operations Research*, Vol. 41, pp. 835–847.

Liu, L., Kashyap, B. R. K., and Templeton, J. G. C. (1990), "On the $GI^X/G/\infty$ System," *Journal of Applied Probability*, Vol. 27, pp. 671–683.

Martin, A. J. (1990), *DRP Distribution Resource Planning: Distribution Management's Most Powerful Tool*, 2nd Ed. Prentice Hall, Englewood Cliffs, NJ, and Oliver Wright, Essex Junction, VT.

Nahmias, S. (1997), *Production and Operations Analysis*, 3rd Ed., Irwin, Chicago.

Press, W. H., Teukolsky, S. A., Vettering, W. T., and Flannery, B. P. (1994), *Numerical Recipes in C*, 2nd Ed., Cambridge University Press, New York.

Ross, S. M. (1996), *Stochastic Processes*, 2nd Ed., John Wiley & Sons, New York.

Scarf, H. (1960), "The Optimality of (s, S) Policies in the Dynamic Inventory Problem," in *Mathematical Methods in the Social Sciences*, K. Arrow, S. Karlin and P. Suppes, Eds., Stanford University Press, Stanford, CA.

Silver, E. A., Pyke, D. F., and Peterson, R. (1998), *Inventory Management and Production Planning and Scheduling*, John Wiley & Sons, New York.

Song, J. S. (1998), "On the Order Fill Rate in a Multi-Item, Base-Stock Inventory System," *Operations Research*, Vol. 46, pp. 831–845.

Song, J. S., Xu, S., and Liu, B. (1999), "Order Fulfillment Performance Measures in an Assemble-to-Order System with Stochastic Leadtimes," *Operations Research*, Vol. 47, pp. 131–149.

Song, J. S., and Yao, D. D. (2000), "Performance Analysis and Optimization of Assemble-to-Order Systems with Random Leadtimes," preprint.

Stenger, A. J. (1994), "Distribution Resource Planning," in *The Logistics Handbook*, J. F. Robeson and W. C. Copacino, Eds., Free Press, New York.

Tijms, H. (1994), *Stochastic Modeling and Analysis: A Computational Approach*, John Wiley & Sons, New York.

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

Yao, D. D., and Zheng, S. (1999), Sequential Inspection under Capacity Constraints. *Operations Research*, Vol. 47, pp. 410–421.

Yao, D. D. and Zheng, S. (1999), ‘‘Coordinated Quality Control in a Two-Stage System,’’ *IEEE Transactions on Automatic Control*, Vol. 44, pp. 1166–1179.

Zipkin, P. (2000), *Foundations of Inventory Management*, Irwin/McGraw-Hill, New York.