Managing Capacity and Inventory Jointly in Manufacturing Systems

James R. Bradley • Peter W. Glynn
Johnson Graduate School of Management, Cornell University, Ithaca, New York 14853-6021
MSE Department, Stanford University, Stanford, California 94305-4026
jrb28@cornell.edu • glynn@leland.stanford.edu

In this paper, we develop approximations that yield insight into the joint optimization of capacity and inventory, and how the optimal inventory policy varies with capacity investment in a single-product, single-station, make-to-stock manufacturing system in which inventory is managed through a base-stock policy. We allow for a correlated demand stream as we analyze our models in an asymptotic regime, in which the penalty and holding costs are small relative to the cost of capacity. Although our approximations are asymptotically correct, our Brownian approximation is accurate even under moderate traffic intensity.
(*Capacity Decisions; Inventory Management; Inventory-Production Policies; Stochastic Modeling; Approximations; Diffusion Processes*)

1. Introduction

Managers are generally aware that capacity and inventory policies must simultaneously be determined for a firm to be optimally managed. However, capacity and inventory decisions are often made separately in organizations either because considering both factors simultaneously is too complex, or because it is assumed that the effect of the interaction between capacity and inventory is small. For example, the hierarchical planning literature suggests that strategic and operational decisions should be made separately: Capacity investment is a strategic decision that should be made by high-level managers; inventory management should be the responsibility of lower-level managers subject to the capacity decisions made by their superiors (Bitran et al. 1981, 1982, Meal 1984).

We consider models with limited capacity, so that we are able to show how the resulting inflexibility affects inventory cost. In this capacitated context, Federgruen and Zipkin (1986a, b) show that a base-stock policy, which we use in all of our analysis, is optimal. Tayur (1993) shows the correspondence between the finite dam problem and capacitated systems, and computes the optimal base-stock policy. Glasserman (1997) develops an expression for the optimal safety stock in a capacitated system as the optimal service level asymptotically approaches 100%. Williams (1984), Zipkin (1986), Buzacott and Shanthikumar (1994), Lambrecht et al. (1984), and Karmarkar and Kekre (1989) also analyze inventory in capacitated production systems. This research however treats the capacity level as fixed, which precludes the optimization of capacity and an analysis of how the optimal inventory policy should change as the capacity level varies.

In this paper, we treat capacity and inventory as joint decision variables, and analyze the interaction of these variables from the perspective of cost. We compute the optimal operating cost as a function of capacity, and also the optimal capacity and inventory policies that minimize long-run average operating cost. By viewing capacity as a decision variable, we extend the work of Federgruen and Zipkin (1986a, b), Glasserman (1997), and Tayur (1993), in which optimal inventory policies are found for capacitated systems with fixed capacities. Our analysis provides for an explicit expression that describes how

0025-1909/02/4802/0273$5.00
1526-5501 electronic ISSN

inventory should optimally be substituted for capacity (to minimize cost) as the capacity level varies. Thus, we provide the first method for scientifically addressing these joint decisions.

We first consider a model in which the order arrivals are governed by a general renewal process, whereas the order processing times are exponentially distributed. We compute the optimal capacity and base-stock levels as the solutions to a pair of nonlinear equations for this model. To obtain analytic insight into the optimal solutions, we consider an asymptotic regime in which the penalty and holding costs are small relative to the capacity cost. Many real-world systems possess this characteristic, for which we are able to develop closed-form approximations to the optimal capacity and base-stock levels (see §2).

In this asymptotic setting, the high cost of capacity forces the firm to set capacity at a level that leads to high utilization at the production facility. This forces the production facility into what is known in queueing theory as a "heavy-traffic" regime, in which the system can be approximated with a Brownian motion model. In §3, we use such Brownian approximations to develop formulae for the optimal decision variables under very weak distributional restrictions on the model inputs (these restrictions allow, for example, correlated arrival processes). In §4, we explore numerically the quality of the approximations developed in this paper and find that the Brownian approximation is accurate even under moderate capacity utilization. We provide some extensions to our basic model in §5 for the circumstances in which delivery times are nonzero, and forecast data guides production decisions. We conclude in §6.

2. The GI/M/1 Manufacturing Model

We consider the joint optimization of the capacity and inventory investment decisions in a manufacturing model with an integrated manufacturing facility and inventory. We assume that demand is stochastic. The objective is to minimize the long-run average operating cost of the manufacturing system.

We use a single-stage, single-server, produce-to-stock manufacturing model in which a single product is produced. Orders are fulfilled from a finished-goods inventory, if the product is available. All unsatisfied demand is backlogged and subsequently fulfilled on a first-come, first-served basis, when the product becomes available. We assume that finished-goods inventory is managed using a base-stock policy, which Federgruen and Zipkin (1986a, b) have shown is optimal for single-stage production systems with limited capacity. The shortfall process, as used by Tayur (1993) and Glasserman (1997), is used to develop the cost function.

We now describe the dynamics of the model. For $n \geq 1$, let A_n be the time at which the nth order is placed so that $U_n = A_n - A_{n-1}$ is the interarrival time of the nth order. Let V_n represent the amount of time required to process the nth order in a unit-rate facility. Thus, in a manufacturing facility having a processing rate β, the time required to process the nth order is V_n/β. The counting process $N_A(t) = \max\{n \geq 0 : A_n \leq t\}$ counts the number of orders placed by time t (where we have adopted the convention that $A_0 = 0$). Also, let $N_V(t) = \max\{n \geq 0 : \sum_{i=1}^{n} V_i \leq t\}$ be the number of units of production completed by a unit-rate facility in its first t units of operational time, so that $N_V(\beta t)$ is the number of units produced in a β-rate facility in its first t units of operational time.

Remark 1. We have explicitly represented the manufacturing capacity in a deterministic fashion: The manufacturing server works at a constant, uninterrupted pace. The rate at which capacity realistically operates, however, is generally stochastic. Such variability can be included in the V_n terms, which could also include work-load variability because of different types of orders.

For a production facility having processing rate β, let $Y_\beta(t)$ be the inventory shortfall process at time t, so that $Y_\beta(\cdot)$ is a nonnegative process that describes the number of units by which the inventory is short of the base-stock level s ($Y_\beta(\cdot)$ is independent of s). By definition of the base-stock policy, the production facility operates if and only if the shortfall is positive. It follows that $Y_\beta = (Y_\beta(t) : t \geq 0)$ satisfies the equation,

$$Y_\beta(t) = Y_\beta(0) + N_A(t) - N_V\left(\beta \int_0^t \mathbf{1}(Y_\beta(u) > 0)\, du\right), \quad (1)$$

where $\mathbf{1}(x) = 1$ if x is true, and zero otherwise. The inventory level process associated with a β-rate production facility is therefore given by

$$I_\beta(s, t) = [s - Y_\beta(t)]^+, \tag{2}$$

where $[x]^+ = x$ if $x > 0$, and zero otherwise. The number of units backordered at time t in a β-rate facility is just

$$\Lambda_\beta(s, t) = [Y_\beta(t) - s]^+. \tag{3}$$

Remark 2. Note from Equation (1) that the dynamics of Y_β are identical to the number-in-system process $Q_\beta = (Q_\beta(t) : t \geq 0)$ associated with a single-server queue with infinite capacity buffer and first-come, first-served queue discipline. In the queueing context, A_n is interpreted as the arrival time of the nth customer and V_n/β is the service time of the nth customer. Because Y_β and Q_β are identical processes, we may apply known results from queueing theory to analyze our system.

As mentioned in the introduction, our analysis of the relationship between capacity and inventory is cost-based. For a time horizon of length t, we will assume that the cost associated with the β-rate system is given by

$$\int_0^t [hI_\beta(s, u) + p\Lambda_\beta(s, u) + c(\beta)]\, du, \tag{4}$$

where h and p are positive constants and $c : [0, \infty) \to [0, \infty)$ is a three-times continuously differentiable strictly increasing function. As is usual in inventory theory modeling, h is the unit holding cost per unit time for inventory and p is the unit backorder penalty cost per unit time. For each choice of $\beta \geq 0$, $c(\beta)$ is the cost per unit time associated with running a facility having such a capacity. Such costs that vary with the capacity level might include amortization costs, certain maintenance costs, and some direct labor costs.

Remark 3. One reasonable "first-cut" choice of $c(\cdot)$ is

$$c(\beta) = c_0\mathbf{1}(\beta > 0) + c\beta, \tag{5}$$

where c_0 and c are positive constants. The first term represents fixed costs of capacity that are incurred regardless of the capacity level selected. The second term represents costs that are incurred proportionally with the capacity level, and regardless of whether the facility is operating or idle. Such costs include the amortization of acquisition cost, certain maintenance costs, and direct labor costs that are often incurred regardless of whether the facility is operating or idle. (For example, some labor contracts require that employees be compensated independently of whether the facility is operating or idle.) Besides providing for a realistic cost structure, the form of this capacity cost function also exhibits a substantial degree of tractability, and so we use (5) in our analysis while omitting the fixed cost given rise to by c_0, which does not influence the optimal capacity and inventory decisions.

Remark 4. Although a capacity function in the form of (5) is appropriate in some industries, such as in the fabrication of integrated circuits (Angelus et al. 1997), one might argue that a concave function $c(\beta)$, which demonstrates economies of scale, is often more realistic. While a capacity function of the form (5) effectively gives significant insight into the capacity-inventory relationship, it also allows for the straightforward analysis for such a concave capacity cost, provided that capacity cost is piecewise linear.

Under very mild assumptions on $(A_n : n \geq 0)$ and $(V_n : n \geq 1)$, it follows that the law of large numbers,

$$\begin{aligned}\frac{1}{t}\int_0^t [hI_\beta(s, u) + p\Lambda_\beta(s, u) + c(\beta)]\, du \\ \to h\mathbf{E}[s - Y_\beta(\infty)]^+ + p\mathbf{E}[Y_\beta(\infty) - s]^+ + c(\beta),\end{aligned} \tag{6}$$

holds as $t \to \infty$ with probability one (Asmussen 1987), where $\mathbf{E}$ denotes expectation. Here, $Y_\beta(\infty)$ is a random variable to which we will refer as the steady-state shortfall random variable associated with a β-rate production facility. As a consequence of the law of large numbers, we find that

$$\begin{aligned}\int_0^t [hI_\beta(s, u) + p\Lambda_\beta(s, u) + c(\beta)]\, du \\ \approx t\{h\mathbf{E}[s - Y_\beta(\infty)]^+ + p\mathbf{E}[Y_\beta(\infty) - s]^+ + c(\beta)\}\end{aligned} \tag{7}$$

for t large. Thus, to minimize (4), we will instead minimize the right-hand side of (6), which we denote by $r(\beta, s)$.

We first analyze a special case of our model that provides the flavor of our results. We assume that orders arrive according to a renewal process, so that $U = (U_n : n \geq 1)$ is a sequence of independent and identically distributed (i.i.d.) random variables with mean λ^{-1}. We further require that the manufacturing times $V = (V_n : n \geq 1)$ are an i.i.d. sequence of exponentially distributed random variables with mean v, and independent of U. Here, Y_β is identical to the number-in-system process for the GI/M/1 queueing system with arrival rate λ and service rate β/v. It is well known that if $\lambda < \beta/v$, then Y_β has a steady-state distribution given by

$$\begin{aligned} P(Y_\beta(\infty) = 0) &= 1 - \rho, \\ P(Y_\beta(\infty) = k) &= \rho(1-\theta)\theta^{k-1}, \end{aligned}$$

for $k \geq 1$ where $\rho = \lambda v/\beta$, $\theta = 1 - \eta v/\beta$, and η is the positive root of the equation $1 - \eta v/\beta = \mathbf{E}e^{-\eta U_1}$ (see, for example, Asmussen 1987, p. 209). Here, $r(\beta, s) = +\infty$ if $\lambda \geq \beta/v$ and

$$\begin{aligned} r(\beta, s) = h\left\{s - \frac{\lambda v}{\beta}[1-\theta(\beta)^s][1-\theta(\beta)]^{-1}\right\} \\ + p\frac{\lambda v}{\beta}\theta(\beta)^s[1-\theta(\beta)]^{-1} + c(\beta) \end{aligned}$$

for $\lambda < \beta/v$. (Here, we write $\theta = \theta(\beta)$ to make clear its dependence on β.) Let $F_\beta(\cdot) = P(Y_\beta(\infty) \leq \cdot)$ be the cumulative distribution function of the steady-state random variable $Y_\beta(\infty)$. Also, define $F_\beta^{-1}(\cdot) = \min_{z\geq 0}\{z : F_\beta(z) \geq \cdot\}$. Then, for $\lambda < \beta/v$ it is easily seen that the increasing nature of $r(\beta, \cdot + 1) - r(\beta, \cdot)$ implies that the optimal base-stock level $s^*(\beta)$ is given by

$$\begin{aligned} s^*(\beta) &= F_\beta^{-1}(p/(p+h)) \\ &= \lceil(\gamma + \ln \rho)/\ln[1/\theta(\beta)]\rceil, \end{aligned} \tag{8}$$

where $\gamma = \ln[(p+h)/h]$. Modulo the restriction on integer values of s, this is a critical fractile solution (as in the newsvendor solution, but with respect to the steady-state shortfall distribution, rather than the demand distribution that applies there). This type of critical fractile solution has appeared in previous shortfall analyses by Tayur (1993) and Wein (1992). Given (8), the function $r(\beta, s^*(\beta))$ can be easily computed numerically over the decision variable β using, for example, easily available spreadsheet optimization programs.

To obtain further insight into the dependence of the optimal base-stock level $s^*(\beta)$ on β, we consider the behavior of the system as the utilization of the facility goes to one (i.e., as $\beta \searrow \lambda v$). To develop a suitable set of approximations, observe that the positive root $\eta \searrow 0$ as $\beta \searrow \lambda v$. Expanding $\mathbf{E}e^{-\eta U_j}$ in a Taylor series about the origin establishes that η must satisfy the quadratic approximation,

$$1 - \frac{\eta v}{\beta} \approx \mathbf{E}\left(1 - \eta U_j + \frac{\eta^2}{2}U_j^2\right),$$

from which we can conclude that

$$\eta \sim \frac{2\lambda}{1+\tau^2}\left(1 - \frac{\lambda v}{\beta}\right),$$

as $\beta \searrow \lambda v$, where $\tau^2 = \operatorname{var} U_n/(\mathbf{E}U_n)^2$ (so that τ^2 is the squared coefficient of variation of the U_js). Here, we use the notation $f(\beta) \sim g(\beta)$ as $\beta \searrow \lambda v$ to denote that $f(\beta)/g(\beta) \to 1$ as $\beta \searrow \lambda v$. Then, it is easily shown that

$$\begin{aligned} s^*(\beta) &\sim \frac{\gamma}{2}(1+\tau^2)\left(1 - \frac{\lambda v}{\beta}\right)^{-1}, \\ \mathbf{E}I_\beta(\infty) &\sim \frac{(1+\tau^2)}{2}[\gamma - (1-e^{-\gamma})]\left(1 - \frac{\lambda v}{\beta}\right)^{-1}, \\ \mathbf{E}\Lambda_\beta(\infty) &\sim \frac{(1+\tau^2)}{2}e^{-\gamma}\left(1 - \frac{\lambda v}{\beta}\right)^{-1}, \end{aligned} \tag{9}$$

as $\beta \searrow \lambda v$. Here, $h\mathbf{E}I_\beta(\infty)$ and $p\mathbf{E}\Lambda_\beta(\infty)$ are, respectively, the expected inventory holding and penalty costs in steady state associated with using the optimal base-stock level for a β-rate system. Thus from (9), and consistent with Glasserman's (1997) conjugate analysis, a manager using the optimal management policy should be prepared to substantially increase the base stock and incur large holding and backorder costs as $\beta \searrow \lambda v$. The optimal base stock increases when capacity decreases to limit the number of backorders in the face of a stochastically greater shortfall distribution. Stated from a sample-path perspective, the base stock must increase as β decreases to limit backorders, because the capacity is less capable of quickly replenishing inventory when demand surges occur.

We conclude the discussion of the GI/M/1 manufacturing model by noting that when the cost of capacity is large relative to the holding and penalty costs, the optimal capacity level β^* is forced to decrease. We introduce here the parameter $\epsilon = \max(p, h)$ to develop an approximation. (We would normally expect $p > h$.)

We shall compute an approximation to our optimal decision variables $(\beta^*, s^*(\beta^*))$ that is valid as $\epsilon \downarrow 0$. We proceed via an informal argument that can be made rigorous mathematically without great difficulty (but with some loss in clarity of exposition).

Note that we must constrain β, $\beta > \lambda v$, so that cost is finite. As $\epsilon \downarrow 0$, $\beta^* \downarrow \lambda v$ so that the optimal capacity cost, for small ϵ, is given approximately by $c(\lambda v) + c'(\lambda v)(\beta^* - \lambda v)$. Relation (9) shows that the sum of expected holding and penalty costs increases in proportion to $\epsilon(\beta^* - \lambda v)^{-1}$ as $\epsilon \downarrow 0$. In an optimal allocation β^* of capacity, the sum of the expected holding and backorder penalty costs should clearly be of the same order of magnitude as the approximate avoidable component of the capacity costs, namely $c'(\lambda v)(\beta^* - \lambda v)$. If we further assume that $c'(\lambda v) > 0$ (capacity cost increases in the capacity level), it follows that $\beta^* = \beta^*(\epsilon)$ takes the form $\beta^*(\epsilon) \approx \lambda v(1 - a^*\epsilon^{\frac{1}{2}})^{-1}$ for ϵ small, where a^* is a positive constant to be determined. Let $r(a) = r(\lambda v(1 - a\epsilon^{\frac{1}{2}})^{-1}, s^*(\lambda v(1-a\epsilon^{\frac{1}{2}})^{-1}))$ be the steady-state expected cost per unit time when the optimal base-stock level is used in conjunction with a capacity of the form $\lambda v(1-a\epsilon^{\frac{1}{2}})^{-1}$. Some routine computations prove that

$$r(a) - c(\lambda v) \sim \epsilon^{\frac{1}{2}}\left\{\frac{(1+\tau^2)}{2a}\left\{\tilde{h}[\gamma - (1-e^{-\gamma})] + \tilde{p}e^{-\gamma}\right\} + c'(\lambda v)a\lambda v\right\} \tag{10}$$

as $\epsilon \downarrow 0$, where $\tilde{h} = h/\epsilon$ and $\tilde{p} = p/\epsilon$. By minimizing the right-hand side of (10) over a, we may conclude that

$$\begin{aligned} \beta^*(\epsilon) &\sim \lambda v\left(1 - a^*\epsilon^{\frac{1}{2}}\right)^{-1}, \\ s^*(\beta^*(\epsilon)) &\sim \frac{\gamma}{2a^*}\epsilon^{-\frac{1}{2}}(1+\tau^2) \end{aligned} \tag{11}$$

as $\epsilon \downarrow 0$, where

$$a^* = \sqrt{\frac{(1+\tau^2)}{2\lambda v c'(\lambda v)}\{\tilde{h}[\gamma - (1-e^{-\gamma})] + \tilde{p}e^{-\gamma}\}}. \tag{12}$$

The expression above assumes that p and h are small, relative to the capacity cost. We can, in fact, define a regime of parameter values for which (11) and (12) provide an appropriate approximation. We, of course, desire that $\beta^*(\epsilon) > 0$. Thus, we require that $a^*\epsilon^{1/2} < 1$. This requires, for example, that $c'(\lambda v)$ be sufficiently large relative to inventory costs, which is the precise circumstance for which we expect this approximation to be valid. From a practical standpoint, such a combination of parameters is often quite reasonable. As we shall see later, the analytic tractability that is associated with this "asymptotic regime" has to do with the fact that a single-server queue in "heavy traffic" can be approximated by an analytically tractable reflected Brownian motion process.

Example 1. One important special case of the above GI/M/1 model is that in which the arrival process is Poisson with rate $\lambda > 0$. In this case, $\tau^2 = 1$ and (11) provides an approximation to the optimal decision variables for the M/M/1 version of our manufacturing model when ϵ is small. Also, $\theta(\beta) = \lambda v/\beta$, which used in (8) gives the optimal base stock.

Example 2. Suppose that the orders arrive at equally spaced time intervals, each of length λ^{-1}. Again, (11) provides an approximation to the optimal decision variables that are valid for ϵ small. Here, $\tau^2 = 0$, and so the approximation of the optimal base-stock level is smaller than in the M/M/1 setting by a factor of $\sqrt{2}$. The positive root η can be found by solving $1 - \eta v/\beta = e^{-\eta/\lambda}$. Thus, $\theta(\beta)$ is determined and the optimal base stock can be found via (8).

3. A Brownian Motion Model

In §2, we were able to make a number of explicit computations for the GI/M/1 model, in which we assumed that the order placement process was i.i.d., and the processing times were exponential. Such assumptions are, of course, unreasonable in many settings. In this section, we provide a Brownian model that offers convenient approximations to the optimal capacity and base-stock levels for the more complex systems that often arise in practice.

Such Brownian approximations are widely used within the queueing community as a means of assessing the performance of complex queueing systems. These queueing approximations typically have provably good performance when the system is in "heavy-traffic" (i.e., when the utilization is close to 1). In our setting, this corresponds to problems in which the holding cost and penalty cost are small as compared to the cost of capacity. As discussed in §2, this parameter regime forces the optimal decision variables to satisfy $\beta/v \approx \lambda$, so that utilization is high. We will see however that the Brownian model provides good guidance for capacity and inventory decisions even when β/v is much greater than λ so that the capacity utilization is much less than one.

To obtain a Brownian approximation to Y_β, we proceed by assuming that Y_β is positive for a large fraction of the time. In view of (1), this suggests that we consider the process

$$\Gamma_\beta(t) = N_A(t) - N_V(\beta t).$$

Our derivation of the Brownian approximation is particularly transparent if we assume that $((U_n, V_n) : n \geq 1)$ is a stationary sequence of random variables. This encompasses the case in which $(U_n : n \geq 1)$ and $(V_n : n \geq 1)$ are i.i.d and independent, but also allows highly complex dependencies among the interarrival and/or processing times. With this assumption in hand, the natural Brownian approximation to $\Gamma_\beta = (\Gamma_\beta(t) : t \geq 0)$ is

$$\Gamma_\beta(t) \overset{\mathcal{D}}{\approx} (\lambda - \beta/v)t + \sigma_\beta B(t),$$

where $\overset{\mathcal{D}}{\approx}$ denotes "has approximately the same distribution as" (and is intended to be purely a heuristic statement without rigorous mathematical meaning) and $B = (B(t) : t \geq 0)$ is a standard Brownian motion (with $\mathbf{E}B(t) = 0$ and $\operatorname{var} B(t) = t$). Here, we choose σ_β^2 to match the time average variance of Γ_β:

$$\sigma_\beta^2 = \lim_{t\to\infty} \frac{1}{t} \operatorname{var}\, \Gamma_\beta(t).$$

But,

$$\begin{aligned}
\Gamma_\beta(t) &- \left(\lambda - \frac{\beta}{v}\right)t \\
&= N_A(t) - \lambda t - \left(N_V(\beta t) - \frac{\beta}{v}t\right) \\
&\approx \lambda\left(\sum_{n=1}^{N_A(t)} (\lambda^{-1} - U_n)\right) - \frac{1}{v}\left(\sum_{n=1}^{N_V(\beta t)} (v - V_n)\right) \\
&\approx \lambda\left(\sum_{n=1}^{\lfloor \lambda t\rfloor} (\lambda^{-1} - U_n)\right) - \frac{1}{v}\left(\sum_{n=1}^{\lfloor \beta t/v\rfloor} (v - V_n)\right). \qquad (13)
\end{aligned}$$

Using the stationarity of $((U_n, V_n) : n \geq 1)$, the variance of (13) can be easily computed. We find that

$$\begin{aligned}
\sigma_\beta^2 = \lambda^3 &\left(\operatorname{var}(U_n) + 2\sum_{j=1}^{\infty} \operatorname{cov}(U_1, U_{j+1})\right) \\
&- \frac{2\lambda^2}{v}\left(\operatorname{cov}(U_1, V_1) + \sum_{j=1}^{\infty}\operatorname{cov}(U_1, V_{j+1}) \right.\\
&\left. + \sum_{j=1}^{\infty}\operatorname{cov}(V_1, U_{j+1})\right) \\
&+ \frac{\beta}{v^3}\left(\operatorname{var}(V_1) + 2\sum_{j=1}^{\infty}\operatorname{cov}(V_1, V_{j+1})\right).
\end{aligned}$$

With the Brownian approximation to Γ_β now computed, we can approximate Y_β by imposing a reflecting barrier in the Brownian motion at the origin. Let $Z_\beta = (Z_\beta(t) : t \geq 0)$ be the corresponding reflecting Brownian motion (RBM) process, so that Z_β is the RBM with drift $\lambda - \beta/v$ and infinitesimal variance σ_β^2, starting at $Y_\beta(0)$. We then approximate the shortfall process Y_β via

$$Y_\beta(\cdot) \overset{\mathcal{D}}{\approx} Z_\beta(\cdot).$$

Because $\lambda < \beta/v$, Z_β has a steady-state $Z_\beta(\infty)$. We therefore approximate $r(\beta, s)$ by

$$r_\beta(\beta, s) = h\mathbf{E}[s - Z_\beta(\infty)]^+ + p\mathbf{E}[Z_\beta(\infty) - s]^+ + c(\beta),$$

which can be expressed in closed form as

$$\begin{aligned}
r_\beta(\beta, s) = hs &- \frac{h\sigma_\beta^2}{2(\frac{\beta}{v} - \lambda)}\left(1 - e^{-2(\beta/v-\lambda)s/\sigma_\beta^2}\right) \\
&+ \frac{p\sigma_\beta^2}{2(\frac{\beta}{v} - \lambda)} e^{-2(\beta/v-\lambda)s/\sigma_\beta^2} + c(\beta).
\end{aligned}$$

For a given capacity β, the optimal base-stock level $s_B^*(\beta)$ is given by

$$s_B^*(\beta) = \frac{\sigma_\beta^2}{2(\frac{\beta}{v} - \lambda)} \ln\left(\frac{p+h}{h}\right), \qquad (14)$$

which we round to the nearest integer in approximating the GI/M/1 and D/G/1 base stock. The function $r_\beta(\beta, s_B^*(\beta))$, given by

$$r_\beta(\beta, s_B^*(\beta)) = \frac{h\sigma_\beta^2}{2(\frac{\beta}{v} - \lambda)} \ln\left(\frac{p+h}{h}\right) + c(\beta), \qquad (15)$$

can then be optimized numerically (closed-form expressions for the optimal capacity level can sometimes be obtained, as is the case with our M/M/1, D/M/1, and D/G/1 models).

For h small, the optimal capacity β^* is forced down to $\beta_0 \overset{\triangle}{=} \lambda v$. If we approximate σ^2_β by $\sigma^2_{\beta_0}$ in (15), which simplifies the capacity optimization, then we obtain

$$r_\beta(\beta) = \frac{h\sigma^2_{\beta_0}}{2(\frac{\beta}{v} - \lambda)} \ln\left(\frac{p+h}{h}\right) + c(\beta). \qquad (16)$$

If we take $c(\cdot)$ of the form $c(\beta) = c(\beta_0) + d(\frac{\beta}{v} - \lambda)^q$ with $d > 0$ and $q > 0$, then we find that the optimal Brownian capacity β^*_B of $r_\beta(\cdot)$ is

$$\beta^*_B = \lambda v + v\left(\frac{h\sigma^2_{\beta_0}}{2dq} \ln\left(\frac{p+h}{h}\right)\right)^{\frac{1}{q+1}}. \qquad (17)$$

We found that using (16) rather than (15) to determine the Brownian capacity did not affect the approximation accuracy; the approximations of the optimal M/M/1, D/M/1, and D/G/1 capacity levels were, in fact, the same for σ^2_β or $\sigma^2_{\beta_0}$. One may alternately use $\sigma^2_{\beta_0}$ in (14) rather than σ^2_β to approximate the base stock. This substitution changes the base-stock approximation, but the quality of the two approaches are comparable. We found that the Brownian base-stock approximation accuracy depends as much on the method one uses to transform the continuous Brownian base-stock parameter into an integer quantity as it does on whether σ^2_β or $\sigma^2_{\beta_0}$ is used in (14). We approximated the optimal capacity with (17), and the optimal base stock with (14), using σ^2_β.

Example 1 (Continued). Recall that here the arrival process is Poisson with rate $\lambda > 0$, and that the processing times are i.i.d. exponential with mean $v > 0$. Here, $\sigma^2_\beta = \lambda + \beta/v$ and $\sigma^2_{\beta_0} = 2\lambda$, and so we obtain the approximations

$$\beta^*_B = \lambda v + v\left(\frac{h\lambda}{dq} \ln\left(\frac{p+h}{h}\right)\right)^{\frac{1}{q+1}},$$
$$s^*_B(\beta^*_B) = \frac{\sigma^2_\beta}{2(\beta^*/v - \lambda)} \ln\left(\frac{p+h}{h}\right).$$

Example 2 (Continued). If the arrivals occur at equally spaced intervals with exponential processing times, then $\sigma^2_\beta = \beta/v$ and $\sigma^2_{\beta_0} = \lambda$, so

$$\beta^*_B = \lambda v + v\left(\frac{h\lambda}{2dq} \ln\left(\frac{p+h}{h}\right)\right)^{\frac{1}{q+1}},$$
$$s^*_B(\beta^*_B) = \frac{\sigma^2_\beta}{2(\beta^*/v - \lambda)} \ln\left(\frac{p+h}{h}\right).$$

Example 3. Suppose that $\operatorname{corr}(U_1, U_{j+1}) = \varrho^j$ with $|\varrho| < 1$, so that the order times are correlated. Assume that $(V_n : n \geq 1)$ is i.i.d. and independent of the order arrival process. In this case,

$$\sigma^2_\beta = \frac{\lambda^3(\operatorname{var} U_1)}{1-\varrho} + \frac{\beta}{v^3}(\operatorname{var} V_1).$$

Thus the optimal capacity level increases relative to the i.i.d. interarrival case when the correlations are positive and decreases when the correlations are negative.

Wein (1992) has discussed the effect of p and h on the optimal base-stock policy. Another interpretation regarding the optimal base stock can be made from Equation (14), which is its equivalence with the mean shortfall, $\sigma^2_{\beta^*}/[2(\beta/v - \lambda)]$, multiplied by $\gamma = \ln[(p+h)/h]$, a factor that is a function of inventory holding and penalty costs. The "safety" factor γ scales the optimal Brownian, and also the optimal GI/M/1 base stock in accordance with the inventory costs. This observation might allow a simple rule for setting the base stock in practice without complex computations: Simply observe the mean shortfall, and set γ according to the inventory cost parameters. Equation (17) shows that the optimal capacity is equal to the mean demand rate λv plus a capacity safety factor that is increasing in h, p, and $\sigma^2_{\beta_0}$, and decreasing in d and q.

The structure of the optimal long-run average cost as a function of capacity for the Brownian model is $r(\beta) = hs^*(\beta) + c(\beta)$. Ignoring the effect of integer demand, this structure also applies to the M/M/1 and D/M/1 models. Inventory costs are simply $hs^*(\beta)$. Wein (1992) found inventory costs to be of this form for an exponential shortfall model (as is the shortfall of our Brownian model) as did Robinson (1993) for a broad class of (Q, r) inventory models.

In the next section, we evaluate the accuracy of the asymptotic closed-form expressions that we developed in §2, and the Brownian approximation that we developed in this section for the optimal GI/M/1 capacity and base stock. Before doing so however, we note that the Brownian model is applicable to a broader class of models than is the asymptotic approximation, which makes a Brownian approximation of a wide variety of manufacturing systems possible, including a D/G/1 model, which we evaluate in the next section also.

4. Approximating the Optimal Capacity and Base Stock

In this section, we analyze the accuracy with which the Brownian model and the GI/M/1 asymptotic expressions approximate the optimal solutions of the M/M/1 and D/M/1 models. We also analyze the Brownian approximation for a D/G/1 model in which interarrival times between orders are constant, $A_n = n/\lambda$ for $n \geq 1$, and the manufacturing times are distributed according to a lognormal distribution. That is, $V = (V_n : n \geq 1)$ are i.i.d. lognormal distributed random variables with mean v. As usual, we require $\beta > \lambda v$. We chose the lognormal distribution for manufacturing times in this model because the independence of the variance and mean allows us to analyze manufacturing times over a range of coefficients of variation, and it is skewed.

We assume that the capacity cost is of the form $c(\beta) = c(\lambda v) + d(\beta/v - \lambda)^q$ for $\beta \geq \lambda v$, and we set $q = 1$ and $c(\lambda v) = d\lambda$ so that $c(\beta) = d\beta/v$ is of the form (5) with $c = d/v$. Throughout this section, and in Tables 3 and 4, we use $r(\cdot)$ to denote the cost function for any system under consideration: either the M/M/1, D/M/1, or D/G/1 models. We denote the optimal solution to $r(\cdot)$ as β^* and $s^*(\beta^*)$. Consistent with previous notation, $r(\beta^*(\epsilon), s^*(\beta^*(\epsilon)))$ and $r(\beta_B^*, s_B^*(\beta_B^*))$ denote the cost of operating a particular system using the asymptotic and Brownian approximations for the optimal capacity and base stock, respectively.

Fifty-four combinations of h, p, and d were analyzed as shown in Table 1. Only these three parameters need be varied, because varying λ and v does not yield any additional information. The ratio of the optimal capacity level and mean unit workload, β^*/v, remains constant for the M/M/1, D/M/1, and D/G/1 models as v varies, while the optimal cost, $r(\beta^*)$, remains unchanged. The same effect holds for the asymptotic and Brownian approximations as v changes, and so we do not need to vary v because the approximation error will remain unchanged. The GI/M/1 models, the D/G/1 model, and the approximations all exhibit similar behavior when λd is held constant. In this case, the ratios of the optimal capacity and the approximated optimal capacities to the arrival rate, β^*/λ, $\beta^*(\epsilon)/\lambda$, and β_B^*/λ, remain constant. Simultaneously, the optimal base-stock parameters and the approximations of the optimal base stock remain constant. Thus, varying both λ and d can result in repetition of equivalent scenarios. Because we can gather an equivalent amount of data by varying just d rather than λ and d, we do so with $\lambda = 1$ and $v = 1$. The values of the D/G/1 model manufacturing time variance that we used are shown in Table 2.

Table 1 **Simulation Cost Parameters**

Parameter	Values
h	1
p	2, 5, 10, 15, 25, 50
d	0.5, 1, 5, 10, 25, 50, 100, 250, 500

The optimal solutions and performance measures for the M/M/1 and D/M/1 models, as noted earlier, can be found numerically using spreadsheet optimization programs and the analysis in §2. The corresponding queueing model is no longer of the GI/M/1 type when the manufacturing time distribution is lognormal, and so we used simulation to approximate the optimal D/G/1 solutions to assess the accuracy of the Brownian approximation for that model.

A comprehensive description of the simulation program, written in C++ to search for the optimal D/G/1 solution, can be found in an expanded version

Table 2 D/G/1 **Manufacturing Time Distribution Parameters**

Scenario	v	var V_1	SCV	$\sigma^2_{\beta_0}$
1	1	0.5	0.5	0.5
2	1	1	1	1
3	1	5	5	5

Note. SCV denotes squared coefficient of variance.

of this paper at the first author's website.[1] In short, the simulation performed a one-dimensional safeguarded Newton search. The operating cost function, $r(\beta, s(\beta))$, in general does not satisfy convexity conditions that guarantee the global optimality of any solution to which the optimization algorithm might converge. Thus, the optimization was written so each local minima would be found if evidence gathered during the search suggested the existence of multiple minima.

Data that compares the Brownian and asymptotic approximation to the optimal solutions for the M/M/1 and D/M/1 models are shown in Table 3 and data documenting the accuracy of the Brownian approximation for the D/G/1 model are shown in Table 4. These tables show only partial results of the simulation study. A complete summary of results can be found in the extended version of this paper.

Our numerical analysis showed, as we would suspect, that the accuracy of both the asymptotic and Brownian approximations increased as d increased and drove the optimal solution toward heavy traffic. For no parameter combination was the asymptotic approximation superior to the Brownian approximation of the M/M/1 and D/M/1 models. The relative error of the Brownian approximation for the M/M/1 and D/M/1 systems was less than 0.13% and 0.64% of the optimal cost, respectively, in cases when the Brownian capacity approximation represented a utilization greater than 80%, and less than 0.81% and 4.96% when the Brownian model suggested an optimal capacity utilization 60% or greater. The greatest relative error of the asymptotic approximation when it suggested capacity utilization greater than 80% was 1.15% for the M/M/1 model, and 1.83% for the D/M/1 model. In only one of the 162 parameter combinations for the D/G/1 model did the relative cost error exceed 2% when the Brownian capacity approximation represented capacity utilization greater than 80%. The efficacy of the Brownian approximation as demonstrated by these results lies not in its capability of predicting the operating cost accurately, but that the optimal solution of the Brownian model is often very close to that of the actual system.

Table 3 Approximations for GI/M/1 **Model** ($\lambda = 1, v = 1$)

		M/M/1 Model			D/M/1 Model		
p	d	ρ_B	% Err.^{1}	% Err.^{2}	ρ_B	% Err.^{1}	% Err.^{2}
2	0.5	0.40	0.138	N/A	0.49	1.321	N/A
2	1	0.49	8.223	N/A	0.57	0.051	75.079
2	5	0.68	0.629	10.540	0.75	0.031	5.298
2	10	0.75	0.305	3.642	0.81	0.110	0.895
2	25	0.83	0.126	0.988	0.87	0.002	0.496
2	100	0.91	0.003	0.113	0.93	0.004	0.031
2	500	0.96	0.000	0.011	0.97	0.000	0.007
10	0.5	0.31	1.550	N/A	0.39	24.497	N/A
10	1	0.39	6.812	N/A	0.48	4.580	N/A
10	5	0.59	1.617	46.370	0.67	4.335	11.942
10	10	0.67	0.125	14.682	0.74	1.965	3.909
10	25	0.76	0.101	3.960	0.82	0.637	0.986
10	100	0.87	0.037	0.420	0.90	0.032	0.183
10	500	0.94	0.003	0.057	0.95	0.011	0.012
25	0.5	0.28	10.249	N/A	0.36	50.245	N/A
25	1	0.36	0.078	N/A	0.44	20.217	N/A
25	5	0.55	1.753	105.329	0.64	1.521	26.821
25	10	0.64	0.033	25.309	0.71	2.890	8.289
25	25	0.73	0.280	5.867	0.80	0.646	1.824
25	100	0.85	0.044	0.801	0.89	0.044	0.362
25	500	0.93	0.001	0.070	0.95	0.008	0.042
50	0.5	0.26	4.171	N/A	0.34	39.896	N/A
50	1	0.34	5.670	N/A	0.42	16.226	N/A
50	5	0.53	0.075	208.506	0.61	4.959	35.221
50	10	0.61	0.097	32.875	0.69	1.559	10.110
50	25	0.72	0.178	8.060	0.78	0.966	3.086
50	100	0.83	0.034	1.149	0.88	0.165	0.426
50	500	0.92	0.000	0.098	0.94	0.011	0.034

Note. ρ_B—λ/β_B^*, capacity utilization of Brownian approximation. N/A—Asymptotic approximation not appropriate for this parameter combination. % Err.^{1}—Approximation errors for Brownian approximation, $100 \times [r(\beta_B^*, s_B^*(\beta_B^*)) - r(\beta^*, s^*(\beta^*))]/r(\beta^*, s^*(\beta^*))$. % Err.^{2}—Approximation errors for asymptotic approximation, $100 \times [r(\beta^*(\epsilon), s^*(\beta^*(\epsilon))) - r(\beta^*, s^*(\beta^*))]/r(\beta^*, s^*(\beta^*))$.

Given the Brownian approximation performance at 80% capacity utilization or greater, Equation (17) implies an accurate Brownian approximation for the models considered here with $c(\beta) = d\beta/v$ when

$$\frac{1}{\lambda}\sqrt{\frac{h\sigma_{\beta_0}^2}{2d}\ln\left(\frac{p+h}{h}\right)} < 0.20. \tag{18}$$

Equation (18) implies that the Brownian approximation is more accurate for h small, p small, λ large, and d large, which was verified by our numerical analysis for all models (allowing for integer effects). In the

[1] ⟨http://www.johnson.cornell.edu/facultybios⟩.

Table 4 Brownian D/G/1 Approximation, Scenarios 1 to 3

		Scenario					
		$v=1$, SCV $=0.5$		$v=1$, SCV $=1$		$v=1$, SCV $=5$	
p	d	ρ_B	% Err.	ρ_B	% Err.	ρ_B	% Err.
2	0.5	0.574	50.357	0.488	0.996	0.299	3.947
2	1	0.656	0.037	0.574	0.019	0.376	14.671
2	5	0.810	1.425	0.751	0.009	0.574	7.679
2	25	0.905	0.004	0.871	0.000	0.751	1.791
2	100	0.950	0.042	0.931	0.009	0.858	0.566
2	500	0.977	0.000	0.968	0.000	0.931	0.065
10	0.5	0.477	0.414	0.392	0.709	0.224	13.268
10	1	0.564	3.349	0.477	3.758	0.290	5.232
10	5	0.743	0.803	0.671	3.096	0.477	3.853
10	25	0.866	0.352	0.820	0.791	0.671	1.635
10	100	0.928	0.046	0.901	0.077	0.803	0.426
10	500	0.967	0.021	0.953	0.026	0.901	0.035
25	0.5	0.439	4.581	0.357	5.832	0.199	4.974
25	1	0.526	12.126	0.439	14.175	0.260	6.474
25	5	0.712	6.512	0.637	2.319	0.439	5.670
25	25	0.847	0.226	0.797	1.768	0.637	3.411
25	100	0.917	0.236	0.887	0.217	0.778	1.282
25	500	0.961	0.028	0.946	0.022	0.887	0.354
50	0.5	0.416	12.557	0.335	7.847	0.184	1.704
50	1	0.502	32.799	0.416	35.981	0.242	7.195
50	5	0.693	16.204	0.615	8.056	0.416	9.284
50	25	0.835	1.848	0.781	2.245	0.615	6.684
50	100	0.910	0.318	0.877	0.613	0.761	3.797
50	500	0.958	0.041	0.941	0.096	0.877	1.171

Note. ρ_B—λ/β_B^*, capacity utilization of Brownian approximation. % Err.—Percentage error of Brownian approximation from optimal D/G/1 cost, $100 \times [r(\beta_B^*, s_B^*(\beta_B^*)) - r(\beta^*, s^*(\beta^*))]/r(\beta^*, s^*(\beta^*))$. SCV—Squared coefficient of variation.

D/G/1 approximation, we observed that an increase in $\sigma_{\beta_0}^2$ did not always cause the approximation accuracy to decrease. The most significant deterioration in approximation accuracy as $\sigma_{\beta_0}^2$ increased was for small values of d, for which the approximation was not accurate at any value of $\sigma_{\beta_0}^2$.

In summary, the asymptotic approximation is accurate when capacity costs are large relative to inventory costs, but the Brownian approximation is superior. The Brownian approximation, in fact, provides an accurate approximation for the M/M/1 and D/M/1 models when capacity utilization is as low as 60%, and for the D/G/1 model with lognormal processing times at 80% utilization. The Brownian model also has the added advantage of tractability and applicability to a wide variety of arrival and processing time scenarios.

5. Extensions to the Basic Model

Although the capacity-inventory interaction is a critical kernel of a manufacturing system, other factors also affect performance and decision tradeoffs. Toward a fuller understanding of a manufacturing system, we extend our model in this section to incorporate two additional factors. In one extension, we incorporate a transportation function between the manufacturing facility and the inventory. In a second extension, we develop a model that can be used to assess how the joint capacity-inventory decision is affected when production decisions are based on a forecast rather than directly on the observation of demand.

5.1. Incorporating Transportation Time

For this subsection, we assume that the delivery time from the manufacturing facility to the inventory location is nonzero and is exogenous to the manufacturing and arrival processes. We use the Brownian model to approximate the dynamics of the system in the presence of such a transportation time, and begin the development of the approximation by recalling that the netput process without transportation, $\Gamma_\beta(t)$, is well approximated by a Brownian process:

$$\Gamma_\beta(t) \overset{\mathscr{D}}{\approx} (\lambda - \beta/v)t + \sigma_\beta B(t).$$

Now consider the netput function when outstanding orders are delivered to the inventory after some exogenous lag,

$$\Gamma_\beta^T(t) = N_A(t) - N_D(\beta t),$$

where the superscript T denotes netput process with transportation lag, and $N_D(\beta t)$ denotes the number of orders that are delivered by time t with a β-rate manufacturing facility. Then,

$$N_D(\beta t) = N_V(\beta t) - N_P(\beta t),$$

where $N_P(t)$ is the number of the orders in the delivery pipeline at time t, so that

$$\Gamma_\beta^T(t) = N_A(t) - N_V(\beta t) + N_P(\beta t).$$

Let T_n be the transportation time required for the nth order, which is independent of U_n, V_n for all n. Note that because the average number of orders in the pipeline over the long term by Little's Law is $\lambda \mathbf{E}T_1$, we can write

$$\begin{aligned}\Gamma_\beta^T(t) &- (\lambda-\beta/v)t - \lambda\mathbf{E}T_1\\ &= (N_A(t)-\lambda t) - (N_V(\beta t)-\beta t/v) + N_P(\beta t) - \lambda \mathbf{E}T_1.\end{aligned}$$

However, with very mild restrictions on the delivery lead times, we find that

$$\lim_{t\to\infty}\frac{1}{t}\ \mathrm{var}(N_P(\beta t)-\lambda\mathbf{E}T_1)=0,$$

because the variance in the number of orders in the pipeline is finite and does not grow with time. Then, using the time-average variance as we did in our previous Brownian approximation, we expect the following to be true:

$$\Gamma_\beta^T(t)-\lambda\mathbf{E}T_1 \overset{\mathcal{D}}{\approx} (\lambda-\beta/v)t+\sigma_\beta B(t). \tag{19}$$

That is, we could approximate the shortfall trajectory as Brownian motion with a boundary at $\lambda\mathbf{E}T_1$, rather than the origin. The drift and variance parameters remain unchanged from the approximation for the case without transportation time.

From (14), (15), and (19), we find that the optimal Brownian capacity remains unchanged, and that the optimal Brownian base stock with transportation time is

$$s_T^* = s_B^* + \lambda\mathbf{E}T_1.$$

While our experiments have shown that this approximation is accurate in heavy traffic, we propose here an improved approximation that takes into account the shortfall variance that is induced by the delivery lag. The basis of our revised approximation is the computation of a Brownian variance parameter that comprehends the effect of the delivery lag as well as the manufacturing facility. Specifically, we compute a Brownian variance parameter by adding together two quantities: the variance parameter that is appropriate in heavy-traffic without delivery lag, and a variance parameter that compensates for the portion of shortfall because of the delivery lag.

To compute a variance parameter for the delivery pipeline, we first recall the well-known approximation of the variance of lead-time demand:

$$\sigma_{LTD}^2 = \mu_L\sigma_D^2 + \mu_D^2\sigma_L^2, \tag{20}$$

where μ_L is the mean lead time, and σ_D^2 is the variance of demand, μ_D is the mean demand, and σ_L^2 is the variance of lead time (the variance of demand over any period is assumed to scale linearly with σ_D^2). Next, assuming that we can model the pipeline inventory process as a RBM, we compute an appropriate Brownian variance parameter. We choose the Brownian variance parameter so that the variance of the stationary RBM distribution is equivalent to σ_{LTD}^2.

If σ^2 is the infinitesimal variance of a RBM with drift $(\lambda-\beta/v)$, then the variance of the stationary distribution is $(\sigma^2)^2/[4(\lambda-\beta/v)^2]$. Equating this quantity with σ_{LTD}^2, we solve for σ_P^2, by which we denote the Brownian variance parameter for the pipeline inventory,

$$\begin{aligned}(\sigma_P^2)^2\big/\big[4(\lambda-\beta/v)^2\big] &= \sigma_{LTD}^2\\ \sigma_P^2 &= 2(\beta/v-\lambda)\sqrt{\mu_L\sigma_D^2+\mu_D^2\sigma_L^2}.\end{aligned}$$

Given that the Brownian model of §3 is accurate in heavy traffic with $T=0$, we want to specify a Brownian variance parameter σ^2 so that $\sigma^2=\sigma_{\beta_0}^2$ for $\mathbf{E}T_1$ small. Conversely, we want $\sigma^2=\sigma_{\beta_0}^2+\sigma_P^2$ when $\mathbf{E}T_1$ is large relative to the delay at the manufacturing server to take into account the variance because of both the manufacturing facility and the delivery pipeline. So, we set σ^2 using a function $\varphi(\cdot)$,

$$\sigma^2=\sigma_{\beta_0}^2+\varphi(\mathbf{E}T_1,\rho_B)\sigma_P^2,$$

which depends on the expected delivery lag $\mathbf{E}T_1$ and the utilization of the manufacturing facility under the Brownian approximation, $\rho_B=\lambda v/\beta_B$. (Note that more complex forms of $\varphi(\cdot)$ are possible that depend on higher moments of the delivery lag, manufacturing time, and demand interarrival distributions.) We want $\varphi(\mathbf{E}T_1,\rho_B)\to 0$ as $\mathbf{E}T_1\to 0$, and $\varphi(\mathbf{E}T_1,\rho_B)\to 1$ as $\rho_B\to 0$. We investigate one possible alternative that fits this criterion: $\varphi(\mathbf{E}T_1,\rho_B)=\mathbf{E}T_1/(\mathbf{E}T_1+\mathbf{E}T_S)$, where $\mathbf{E}T_S$ is the expected delay in the manufacturing step alone, which depends on ρ_B.

We investigated the accuracy of this approximation for both (constant) deterministic and uniformly distributed transportation times with exponentially distributed interarrival and manufacturing times. Specifically, we experimented with three cases for deterministic delivery times, $T = 1, 2, 5$, and three cases of delivery lags distributed according to the uniform distribution with corresponding means, $T = U[0, 2], U[0, 4], U[0, 10]$.

We observe in Tables 5 and 6 how the percentage error from optimal cost varies as $\mathbf{E}T_1$ increases as a percentage of the total time in manufacture and transportation, that is, as $\varphi(\mathbf{E}T_1, \rho_B)$ increases toward 1. (We find that $\mathbf{E}T_S = v/(\beta - \lambda v)$ with exponentially distributed interarrival and manufacturing times.) The capacity-base stock approximation is accurate when a substantial portion of the replenishment time is due to transportation; the error is less than 3% of the optimal cost whenever $\varphi(\mathbf{E}T_1, \rho_B) \leq 50\%$. The approximation error is less than 3% for relatively low capacity utilization, for example, when $\rho_B \geq 50\%$ for $\mathbf{E}T_1 = 1$, $\rho_B \geq 60\%$ for $\mathbf{E}T_1 = 2$, and $\rho_B \geq 75\%$ for $\mathbf{E}T_1 = 5$.

Table 5 **Capacity-Inventory Model with Deterministic Transportation Time**

		Scenario								
		$T_1 = 1$			$T_1 = 2$			$T_1 = 5$		
p	c	ρ_B	$\varphi(\cdot)$	% Err.	ρ_B	$\varphi(\cdot)$	% Err.	ρ_B	$\varphi(\cdot)$	% Err.
2	0.5	0.49	0.51	4.21	0.52	0.65	0.00	0.53	0.82	4.22
2	1	0.55	0.45	0.00	0.58	0.59	0.00	0.61	0.76	1.39
2	5	0.69	0.31	0.00	0.71	0.45	0.00	0.73	0.65	0.00
2	10	0.76	0.24	0.10	0.76	0.38	0.00	0.78	0.59	0.06
2	25	0.83	0.17	0.06	0.83	0.29	0.00	0.84	0.50	0.00
2	100	0.91	0.09	0.01	0.91	0.17	0.07	0.91	0.34	0.00
2	500	0.96	0.04	0.03	0.96	0.09	0.03	0.96	0.19	0.08
10	0.5	0.42	0.58	0.03	0.43	0.72	2.63	0.39	0.89	14.23
10	1	0.48	0.52	3.90	0.51	0.66	5.31	0.52	0.82	13.00
10	5	0.62	0.38	0.00	0.64	0.52	1.18	0.68	0.70	2.78
10	10	0.69	0.31	0.00	0.70	0.46	0.00	0.73	0.65	1.11
10	25	0.77	0.23	0.19	0.77	0.37	0.00	0.79	0.57	0.00
10	100	0.87	0.13	0.07	0.87	0.23	0.16	0.87	0.43	0.00
10	500	0.94	0.06	0.08	0.94	0.12	0.05	0.94	0.26	0.04
25	0.5	0.40	0.60	0.00	0.39	0.76	16.63	0.34	0.91	36.85
25	1	0.46	0.54	4.24	0.48	0.68	5.02	0.46	0.85	21.88
25	5	0.60	0.40	0.00	0.62	0.55	0.63	0.65	0.72	4.10
25	10	0.66	0.34	0.41	0.68	0.49	0.00	0.71	0.68	1.69
25	25	0.74	0.26	0.15	0.80	0.33	0.04	0.77	0.60	0.37
25	100	0.85	0.15	0.23	0.85	0.26	0.12	0.85	0.46	0.08
25	500	0.93	0.07	0.21	0.93	0.14	0.00	0.93	0.29	0.00
50	0.5	0.38	0.62	4.38	0.36	0.78	21.00	0.31	0.92	37.06
50	1	0.44	0.56	5.22	0.46	0.70	9.70	0.43	0.87	32.64
50	5	0.58	0.42	0.00	0.61	0.57	3.13	0.64	0.74	7.42
50	10	0.64	0.36	0.20	0.66	0.51	0.08	0.69	0.69	2.62
50	50	0.72	0.28	0.19	0.73	0.42	0.24	0.80	0.56	0.26
50	100	0.84	0.16	0.34	0.84	0.28	0.14	0.84	0.48	0.11
50	500	0.92	0.08	0.08	0.92	0.15	0.00	0.92	0.31	0.13

Note. ρ_B—λ/β_B^*, capacity utilization of Brownian approximation. % Err.—Percentage cost error of Brownian capacity-inventory policy from optimal cost as determined by simulation optimization. % Err. $=$ 0.00 indicates no statistically significant difference between the Brownian approximation and the solution to which the simulation converged given the accuracy of the simulation.

5.2. The Capacity-Inventory Decision with Forecast Updates

Heath and Jackson (1994) and Graves et al. (1986) developed the Martingale Model of Forecast Evolution (MMFE), which characterizes the evolution of forecasts over time. Toktay and Wein (1999) showed how inventory cost in a production-inventory system can be reduced when production decisions are based on forecasts that evolve according to the MMFE by using a base-stock policy that is based on the "forecast-corrected" inventory level, which is the inventory level minus the forecasted demand over a finite horizon. This type of inventory-production policy results in a reduction in inventory cost, although it has not been shown to be optimal in the case of stochastic and limited capacity in the infinite-horizon problem. Our goal in this section is to show how Toktay and Wein's (1999) model can be extended to the problem of jointly optimizing capacity and inventory decisions.

To develop our model, we review the relevant notation and details of the MMFE and Toktay-Wein models. For brevity, we do not review the complete details of these models, which can be found in the aforementioned papers. Note that, for clarity, we do not incorporate the corrected diffusion approximation of Glasserman and Liu (1997) and Siegmund (1979) as did Toktay and Wein (1999). Because the MMFE and Toktay and Wein's model make sense only if decisions and forecasts are made periodically, we adopt a periodic model structure in this section rather than the continuous-time model used previously in this paper.

Table 6 Capacity-Inventory Model with Uniform Transportation Time

		Scenario								
		$T_1 = U\,[0,2]$			$T_1 = U\,[0,4]$			$T_1 = U\,[0,10]$		
p	c	ρ_B	$\varphi(\cdot)$	% Err.	ρ_B	$\varphi(\cdot)$	% Err.	ρ_B	$\varphi(\cdot)$	% Err.
2	0.5	0.50	0.50	2.71	0.59	0.58	4.56	0.72	0.66	19.50
2	1	0.55	0.45	0.00	0.61	0.56	2.74	0.72	0.65	11.66
2	5	0.69	0.31	0.02	0.71	0.45	0.02	0.76	0.61	1.11
2	10	0.75	0.25	0.35	0.76	0.39	0.00	0.79	0.57	0.00
2	25	0.83	0.17	0.00	0.83	0.29	0.12	0.84	0.49	0.00
2	100	0.91	0.09	0.04	0.91	0.17	0.05	0.91	0.34	0.00
2	500	0.96	0.04	0.00	0.96	0.09	0.05	0.96	0.19	0.00
10	0.5	0.45	0.55	1.06	0.58	0.59	12.48	0.72	0.66	38.92
10	1	0.49	0.51	2.54	0.59	0.58	7.81	0.72	0.66	30.56
10	5	0.62	0.38	1.63	0.65	0.51	0.19	0.74	0.64	6.28
10	10	0.68	0.32	1.26	0.70	0.46	0.58	0.76	0.62	2.78
10	25	0.77	0.23	0.08	0.77	0.37	0.33	0.79	0.56	0.00
10	100	0.87	0.13	0.12	0.87	0.24	0.00	0.87	0.43	0.04
10	500	0.94	0.06	0.03	0.94	0.12	0.04	0.94	0.26	0.07
25	0.5	0.44	0.56	1.59	0.57	0.60	19.41	0.72	0.66	59.93
25	1	0.47	0.53	4.35	0.58	0.59	13.76	0.72	0.66	44.61
25	5	0.59	0.41	2.54	0.64	0.53	2.02	0.73	0.65	14.69
25	10	0.65	0.35	0.04	0.68	0.49	0.80	0.75	0.63	4.89
25	25	0.74	0.26	0.78	0.75	0.40	0.08	0.78	0.59	0.24
25	100	0.85	0.15	0.24	0.85	0.26	0.15	0.85	0.46	0.00
25	500	0.93	0.07	0.00	0.93	0.14	0.07	0.93	0.29	0.03
50	0.5	0.43	0.57	0.00	0.57	0.60	20.88	0.72	0.66	65.04
50	1	0.46	0.54	1.73	0.58	0.59	15.73	0.72	0.66	52.90
50	5	0.57	0.43	2.29	0.63	0.54	1.99	0.73	0.65	17.21
50	10	0.64	0.36	2.68	0.67	0.50	1.50	0.74	0.64	7.71
50	25	0.72	0.28	0.00	0.79	0.35	0.25	0.77	0.60	1.30
50	100	0.84	0.16	0.00	0.84	0.28	0.30	0.84	0.48	0.00
50	500	0.92	0.08	0.02	0.92	0.15	0.03	0.92	0.31	0.10

Note. ρ_B—λ/β_B^*, capacity utilization of Brownian approximation. % Err.—Percentage cost error of Brownian capacity-inventory policy from optimal cost as determined by simulation optimization. % Err. = 0.00 indicates no statistically significant difference between the Brownian approximation and the solution to which the simulation converged given the accuracy of the simulation.

Consistent with the remainder of this paper, we consider an integer-valued state space rather than a continuous state space as did Toktay and Wein (1999).

We assume the same sequence of events in each period as did Toktay and Wein (1999). First, demand is revealed, which drives the forecast update. Next, "production authorization" for some number of units is given to the manufacturing facility, after which, production takes place, followed finally by an accounting of inventory costs.

Let D_n be the demand in period n where the demand process $D = (D_n : n \geq 0)$ is stationary with mean $\mathbf{E}D_1 = \lambda$. Let $D_{n,n+i}$ be the forecast of demand in period $n+i$ made in period n. Given the sequence of events, $D_n = D_{n,n}$. It is assumed that meaningful forecasts are made only for H periods in advance, so that $D_{n,n+i} = \lambda$ for $i > H$. The forecast update in period n for demand in period $n+i$ is $\epsilon_{n,n+i} = D_{n,n+i} - D_{n-1,n+i}$. Then define the vector of all nontrivial forecast updates as $\epsilon_n = \{\epsilon_{n,n}, \epsilon_{n,n+1}, \ldots, \epsilon_{n,n+H}\}$. Given Heath and Jackson's (1994) assumptions, ϵ_n for $n \geq 0$ is an i.i.d. multidimensional, normally, distributed random vector with mean zero and covariance matrix Σ. We will denote the elements of Σ as σ_{ij} for $i, j = 0, \ldots, H$. Toktay and Wein (1999) point out that the autocovariance can be computed from the elements of Σ:

$$\gamma_i = \operatorname{cov}(D_n, D_{n+i}) = \sum_{j=0}^{H-i} \sigma_{j,\,j+i} \quad \text{for } i = 0, 1, \ldots, H.$$

The production-inventory system is managed using a "production authorization" system. (See Buzacott and Shanthikumar 1993; using a traditional base-stock or kanban system is such an authorization mechanism with a constant number of "production authorization cards.") Let Q_n be the number of units authorized to be produced, but not yet produced, at the end of period n, and let R_n be the additional number of units that are authorized for production in period n.

Consistent with Toktay and Wein (1999), and in a departure from our foregoing notation, let the production capacity in period n be limited by a random quantity C_n, with mean $\mathbf{E}C_1 = \beta$ and variance σ_C^2. Then the production quantity, P_n, is limited by the minimum of the capacity or the authorized maximum production:

$$P_n = \min(C_n, Q_{n-1} + R_n), \tag{21}$$

and the remaining units authorized for production at the end of the period is

$$Q_n = Q_{n-1} + R_n - P_n. \tag{22}$$

Thus, the inventory level at the end of period n evolves according to

$$I_n = I_{n-1} + P_n - D_n. \tag{23}$$

Toktay and Wein (1999) define the forecast-corrected inventory level to be the inventory level minus the total demand forecasted over the forecast horizon, $\widetilde{I}_n = I_n - \sum_{i=1}^{H} D_{n,\,n+i}$, and show that a base-stock policy with respect to this forecast-corrected inventory level minimizes cost over a finite horizon when capacity is deterministic. The performance of this policy structure in that setting and its simplicity motivates its use in the case of stochastic and limited capacity in the infinite-horizon problem. This policy is implemented using the production authorization quantities

$$R_n = \lambda + \sum_{i=0}^{H} \epsilon_{n,\,n+i}, \tag{24}$$

so that the number of new units authorized for production is equal to the demand in period n plus the cumulative forecast update for the H-period forecasting horizon. Under this policy, it can be shown that the number of units authorized for production at the end of each period plus the forecast-corrected inventory is a constant quantity, which is referred to as the forecast-corrected base-stock level s_H:

$$Q_n + \widetilde{I}_n = s_H \quad \text{for } n = 1, 2, \ldots .$$

Note that the sum of the inventory level I_n plus the number of authorized production units is not constant in this model as is the case with standard base-stock policies. In that sense, guiding the production decision with a forecast causes the base stock, or number of kanban, to be dynamically adjusted according to the expectation of demand.

In a manner analogous with our continuous-time model, let h and p denote the holding and backorder penalty costs of inventory per unit, per period. If $I_\beta(s_H, n)$ and $\Lambda_\beta(s_H, n)$ denote the on-hand inventory and backorder levels for a capacity level β and base-stock level s_H when production is managed according to (21), (22), (23), and (24), then Toktay and Wein's (1999) objective is to minimize inventory cost for a specific capacity level: $hI_\beta(s_H, n) + p\Lambda_\beta(s_H, n)$.

We also are interested in jointly optimizing over the capacity level, and so we consider the following objective, assuming that capacity cost is linear in the capacity level:

$$r(\beta, s_H) = hI_\beta(s_H, n) + p\Lambda_\beta(s_H, n) + c\beta.$$

Toktay and Wein (1999) develop solutions for the optimal base stock in heavy-traffic, where

$$\nu = 2(\beta - \lambda) \Big/ \left(\gamma_0 + 2\sum_{i=1}^{H} \gamma_i + \sigma_C^2 \right)$$

plays a central role as the parameter of the stationary distribution of shortfall.

If $H = 0$ (no forecasts are made), then $R_n = D_n$, Q_n is the shortfall for the periodic model analogous to the shortfall that we have defined for the continuous-time model, and ν is analogous with our exponential parameter. In that case, the optimal base-stock level for the Toktay and Wein (1999) model is also analogous with our previous results:

$$s_0^* = F_{Q_\infty}^{-1}(p/(p+h)),$$

where in heavy traffic $s_0^* = \ln((p+h)/h)/\nu$.

If $H > 0$, then the optimal base-stock policy is $s_H^* = F_W^{-1}(p/(p+h))$, where

$$W = \max\left\{ Q_\infty + X_0, \max_{1 \le k \le H} X_k \right\},$$

and

$$X_k = -k\lambda + \sum_{i=k+1}^{H} \sum_{j=i}^{H} \epsilon_{n-H+i,\, n-H+j} - \sum_{i=1}^{k} \sum_{j=H+1}^{H+i} \epsilon_{n-H+i,\, n-H+j} - \sum_{i=k+1}^{H} C_{n-H+i}.$$

When $p \gg h$, Toktay and Wein (1999) show that $s_H^* \approx s_0^* + \mathbf{E}X_0 + \frac{1}{2}\nu \operatorname{var} X_0$ and that

$$hI_\beta(s_H, n) + p\Lambda_\beta(s_H, n) \approx h(s_H^* + 1/\nu - \mathbf{E}W). \tag{25}$$

It is well known how the covariance of forecasts affects the optimal base-stock level at a given capacity level. Our primary goal upon embarking on this analysis was to use (25) to approximate the optimal (minimum) $r(\beta, s_H^*)$ over β so that we may gain insight into how using an H-period forecast update to guide production decisions affects the optimal capacity level, which we denote by β_H^*. In particular, we are interested in comparing β_H^* with the optimal capacity when the forecast is ignored, in which case $H = 0$ and the optimal capacity is approximated by our Brownian model, β_B^*. We conducted an exhaustive search, in capacity increments of 0.01, using a spreadsheet to find $r(\beta_H^*, s_H^*)$ and $r(\beta_B^*, s_B^*)$ because $r(\beta, s_H^*)$ is in general not convex in β.

We considered the circumstance in which both the number of units demanded and produced in a period are Poisson distributed, with $\lambda = 1$. We also assumed, as did Toktay and Wein (1999) in an example, that demand follows a moving average process with lag 1 such that $D_n = \lambda + e_t - \theta e_{t-1}$, where the demand correlation between demands with lag 1 is $-\theta$. Also, following Toktay and Wein (1999), set $H = 1$, $D_{n,n+1} = \lambda - \theta e_t$, and $D_{n,n+i} = \lambda$ for $i > 1$.

Toktay and Wein (1999) found in a numerical example (which used normal demand rather than our Poisson demand) that approximately 90% capacity utilization was required for (25) to be a reasonably accurate approximation of actual cost (accuracy also depended on p). Thus, in using (25) to compare the optimal capacity strategies for $H = 0$ and $H = 1$, a relatively high capacity utilization is required to ensure that the difference $r(\beta_B^*, s_B^*) - r(\beta_H^*, s_H^*)$ is because of a true difference in cost, rather than approximation error. We performed some exploratory analysis, choosing our parameters carefully to ensure that the points we analyzed fell within the regime where the approximations were accurate. While holding $h = 1$, we evaluated this model for $c = 1, 5, 25, 100, 500, 1{,}000$, and $\theta = -0.90, -0.70, -0.50, -0.30, -0.10, 0.10, 0.30, 0.50, 0.70, 0.90$.

For each combination of h, c, and θ, we then found the smallest value of p such that $\lambda/\beta_H^* \geq 0.90$, if such a value of p existed. We found such values of p for 35 of the 60 parameter combinations. In only one of these cases was $\beta_H^* \neq \beta_B^*$, and in that case $\beta_H^* - \beta_B^* = 0.01$, and $(r(\beta_B^*, s_B^*) - r(\beta_H^*, s_H^*))/r(\beta_H^*, s_H^*) = 0.2\%$. We also found in these circumstances that the base-stock policies were approximately the same, that is, $s_H^* + \lambda \approx s_B^*$. Thus, for relatively high capacity utilization, the optimal base-stock and capacity decisions are equivalent whether a naïve approach is taken or forecasts are used. Further research, which requires more accurate approximations or simulation, is needed to compare β_H^* and β_B^* in circumstances where utilization rates less than 0.90 are optimal.

6. Conclusions

Capacity and inventory decisions are often made separately in practice because either the joint decision is complex or it is assumed that the interaction between the two production factors is small. In the former case, managers might employ a hierarchical decision process, which separates the capacity and inventory decisions, and use rules of thumb to set the capacity level. These rules of thumb sometimes favor high capacity utilization (one might even observe a goal of 100% utilization, see Colgan 1995 and Bradley and Arntzen 1999). One often hears in practice that high capacity utilization is justified because capacity is expensive. Hayes and Wheelwright (1984) support this notion by stating "Unused capacity generally is expensive." We must realize though that total cost of operation is the cost that matters and the cost of capacity should be judged relative to the alternatives. Additional inventory, as we have shown, is required for minimum cost operation when capacity levels are reduced, and capacity may not always be so expensive that high capacity utilization is warranted. In fact, it is obvious from our models that overly restricting the capacity level causes total cost to increase toward infinity as $\beta \searrow \lambda v$ (see (14) and (15) for example).

Our models and approximations offer insight into the capacity-inventory trade-off, and enable better capacity and inventory decisions. The exact base-stock solution, and asymptotic capacity and base-stock approximations for the GI/M/1 model contribute toward this insight. The Brownian model is perhaps more useful though, because it yields closed-form solutions and is applicable to a greater variety of processing time distributions. Moreover, the Brownian model was accurate for the models we tested even when the capacity utilization was significantly less than one. We extended the Brownian approximation to the case of nonzero transportation times, and also performed a preliminary analysis of a manufacturing system in which production decisions are made using forecast data rather than actual demand. Both of these topics warrant further research. Another possible extension of the Brownian model is for the case of nonstationary demand, for example, when demand is seasonal and capacity is constrained below the peak demand.

References

Angelus, A., E. L. Porteus, S. C. Wood. 1997. Optimal sizing and timing of capacity expansion with implications for modular semiconductor wafer fabs. Working paper, Strategic Decisions Group, Menlo Park, CA.

Asmussen, S. 1987. *Applied Probability and Queues*. John Wiley & Sons Ltd., New York.

Bitran, G. R., E. A. Haas, A. C. Hax. 1981. Hierarchical production planning: A single stage system. *Oper. Res.* **29**(4) 717–743.

——, ——, ——. 1982. Hierarchical production planning: A two-stage system. *Oper. Res.* **30**(2) 232–251.

Bradley, J. R., B. C. Arntzen. 1999. The simultaneous planning of production, capacity and inventory in seasonal demand environments. *Oper. Res.* **47**(6) 795–806.

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

——, ——. 1994. Safety stock versus safety time in MRP controlled production systems. *Management Sci.* **40**(12) 1678–1689.

Colgan, J. A. 1995. A business planning model to access the tradeoff between inventory and capacity for a stage 1 manufacturing process. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

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

——, ——. 1986b. An inventory model with limited production capacity and uncertain demand, II: The discounted cost criterion. *Math. Oper. Res.* **11** 208–215.

Glasserman, P. 1997. Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* **45**(2) 244–257.

——, T. Liu. 1997. Corrected diffusion approximations for a multistage production-inventory system. *Math. Oper. Res.* **22**(1) 186–201.

Graves, S. C. 1986. A tactical planning model for a job shop. *Oper. Res.* **34**(4) 522–533.

Harris, F. W. 1913. How many parts to make at once. *Factory, Magazine Management* **2**(10) 135–136, 152.

Hayes, R. H., S. C. Wheelwright. 1984. *Restoring Our Competitive Edge: Competing Through Manufacturing*. John Wiley & Sons, New York.

Heath, D. C., P. L. Jackson. 1994. Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Trans.* **26**(3) 17–30.

Karmarkar, U., S. Kekre. 1989. Batching policy in kanban systems. *J. Manufacturing Systems* **8**(4) 317–328.

Lambrecht, M. R., J. A. Muckstadt, R. Luyten. 1984. Protective Stocks in Multi-stage Production Systems. *Internat. J. Production Res.* **22**(6) 1001–1025.

Meal, H. C. 1984. Putting production decisions where they belong. *Harvard Bus. Rev.* **62**(March–April) 102–111.

Robinson, L. R. 1993. The cost of following the optimal inventory policy. *IIE Trans.* **25**(5) 105–108.

Siegmund, D. 1979. Corrected diffusion approximations for certain random walk problems. *Adv. Appl. Probab.* **11** 701–719.

Tayur, S. 1993. Computing the optimal policy for capacitated inventory models. *Stochastic Models* **9**(4) 585–598.

Toktay, L. B., L. M. Wein. 1999. Analysis of a forecasting-production-inventory system with stationary demand. Working paper, INSEAD, France.

Wein, L. M. 1992. Dynamic scheduling of a multiclass make-to-stock queue. *Oper. Res.* **40**(4) 724–735.

Williams, T. M. 1984. Special products and uncertainty in production/inventory systems. *Eur. J. Oper. Res.* **15** 46–54.

Zipkin, P. H. 1986. Models for design and control of stochastic, multi-item production systems. *Oper. Res.* **34**(1) 91–104.

Accepted by Paul Glasserman; received January 27, 2000. This paper was with the authors 5 months and 17 days for 2 revisions.