Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Stock Positioning and Performance Estimation in Serial Production-Transportation Systems

Guillermo Gallego, Paul Zipkin,

Stock Positioning and Performance Estimation in Serial Production-Transportation Systems

Guillermo Gallego • Paul Zipkin
Department of Industrial Engineering & Operations Research, Columbia University, New York, New York 10027
The Fuqua School of Business, Duke University, Durham, North Carolina 27708

This paper considers serial production-transportation systems. In recent years, researchers have developed a fairly simple functional equation that characterizes optimal system behavior, under the assumption of constant leadtimes. We show that the equation covers a variety of stochastic-leadtime systems as well. Still, many basic managerial issues remain obscure: When should stock be held at upstream stages? Which system attributes drive overall performance, and how? To address these questions, we develop and analyze several heuristic methods, inspired by observation of common practice and numerical experiments. One of these heuristics yields a bound on the optimal average cost. We also study a set of numerical examples, to gain insight into the nature of the optimal solution and to evaluate the heuristics.
(*Inventory/Production*; *Multistage*; *Solutions and Heuristics*)

1. Introduction

Consider a serial production-transportation system.

- There are several *stages*, or stocking points, arranged in series. The first stage receives supplies from an external source. Demand occurs only at the last stage. Demands that cannot be filled immediately are backlogged.
- There is *one product*, or more precisely, one per stage.
- To move units to a stage from its predecessor, the goods must pass through a *supply system*, representing production or transportation activities. The cost for a shipment to each stage is linear in the shipment quantity.
- There is an inventory-holding cost at each stage and a backorder-penalty cost at the last stage. The horizon is infinite, all data are stationary, and the objective is to minimize total average cost. Information and control are centralized.

We focus on a basic system, where time is continuous, demand is a Poisson process, and each stage's supply system generates a constant leadtime. However, virtually all the results remain valid for a discrete-time system with i.i.d. demands, for compound-Poisson demand in continuous time, and for more complex supply systems with stochastic leadtimes. Also, since an assembly system can be reduced to an equivalent series system (Rosling 1989), the results apply there too.

Clark and Scarf (1960) initiated the analysis of this system, assuming discrete time with a finite horizon and nonstationary data. They showed that the optimal policy has a simple, structured form (an echelon base-stock policy) and developed a tractable scheme to compute it. Federgruen and Zipkin (1984) adapted the results to the stationary, infinite-horizon setting and pointed out that the algorithm becomes simpler there. Rosling (1989) and Langenhoff and Zijm (1990) provided streamlined statements of the results. Chen and Zheng (1994) further streamlined the results and extended them to continuous time. The outcome is a fairly simple functional equation, Equation (5) in §2,

1523-4614/99/0101/0077$05.00

that characterizes the optimal policy. See Federgruen (1993) for a review of this literature.

There is another, very different stream of research on multistage systems, one that emphasizes policy evaluation. It assumes a particular policy type, usually a base-stock policy, and estimates key performance measures, especially average inventories and backorders. Those measures are used to construct an optimization model, whose solution yields the best such policy. The supply systems can be fairly complex, indeed some generate stochastic leadtimes. In most cases the performance estimates are approximations. The system structure too can be more complex; in addition to series systems, the approach applies to distribution and assembly systems. This literature begins with the METRIC model of Sherbrooke (1968). Recent contributions include Graves (1985), Sherbrooke (1986), and Svoronos and Zipkin (1991). Reviews can be found in Nahmias (1981) and Axsäter (1993). We explain in §3 that, despite these differences, the solution to Equation (5) also yields the best base-stock policy for such a system, up to the approximation.

Still, many basic managerial issues concerning such systems remain obscure: When should stock be held at upstream stages? Which system attributes drive overall performance, and how? To address these questions, we study several heuristic methods (§4), inspired by observation of common practice and numerical experiments, including one that yields a bound on the optimal average cost. Sensitivity analysis of this result reveals interesting features of system behavior. We also study a set of numerical examples (§5), both to gain insight into the nature of the optimal solution and to evaluate the heuristics.

Section 6 presents our conclusions. A key finding is that system performance is fairly insensitive to stock positioning, provided the overall system inventory is near optimal. In particular, certain heuristic policies which concentrate stock at a few locations perform quite well.

We also discuss a broader system *design* problem, as in Gross et al. (1981). Here, the stages are *potential* storage locations, but none have yet been built. The design problem is to select a subset among them and then to determine a control policy for the resulting network. There is a cost to open each facility, and such costs appear in either the objective function or a constraint. There may be several products sharing the same facilities. This is a hard problem, but several of our heuristics apply to it as well.

2. Base-Stock Policy Evaluation and Optimization

This section reviews the basic facts concerning policy evaluation and optimization.

2.1. Stages

For now, assume Poisson demand and constant leadtimes. Denote

J = number of stages
j = stage index, $j = 1, \dots, J$
λ = demand rate
L_j = supply leadtime to stage j
L = total system leadtime $= \Sigma_j L_j$.

The numbering of stages follows the flow of goods; stage 1 is the first, and stage J is the last, where demand occurs. The external source, which supplies stage 1, has ample stock; it responds immediately to orders.

2.2. Base-Stock Policies

In a single-stage system, a *base-stock policy* aims to keep the inventory position constant. The target inventory position is a policy variable, the *base-stock level*, denoted s. When the inventory position falls below s, the policy orders enough to raise the inventory position to s; otherwise, it does not order. Thus, once the inventory position hits s, orders precisely equal demands.

In a multi-stage system, there are two classes of base-stock policy, local and echelon. Although they seem quite different, the two classes are equivalent (Axsäter and Rosling 1993).

A *local base-stock policy* is a decentralized control scheme, where each stage monitors its own local inventory position and places orders with its predecessor. Each stage j follows a standard, single-stage base-stock policy with parameter

s_j' = local base-stock level for stage j,

a nonnegative integer. The overall policy is characterized by the vector $\mathbf{s}' = (s_j')_{j=1}^{J}$.

An *echelon base-stock policy* is a centralized control scheme. It monitors each stage's echelon inventory (the stage's own stock and everything downstream), and determines external orders and inter-stage shipments according to a base-stock policy. The policy parameters are

s_j = echelon base-stock level for stage j,

also a nonnegative integer. Let $\mathbf{s} = (s_j)_{j=1}^J$. As shown by Chen and Zheng, given stationary parameters, such a policy is optimal in either a periodic-review or a continuous-review setting.

Given a local base-stock policy $\mathbf{s}'$, an equivalent echelon base-stock policy has parameters $s_j = \Sigma_{i \geq j} s_i'$. Conversely, starting with an echelon base-stock policy $\mathbf{s}$, one can construct an equivalent local policy, setting $s_j^- = \min_{i \leq j} \{s_i\}$ and $s_j' = s_j^- - s_{j+1}^-$, where $s_{J+1}^- = 0$. (Also, the echelon base-stock policy $\mathbf{s}^- = (s_j^-)_{j=1}^J$ is equivalent to $\mathbf{s}$.)

2.3. Cost

Denote

$E[\cdot]$ = expectation
$[x]^+$ = $\max\{0, x\}$
$D(t)$ = cumulative demand in the interval $(0, t]$.
$V[\cdot]$ = variance
$[x]^-$ = $\max\{0, -x\}$

The following are state random variables in equilibrium:

I_j' = local on-hand inventory at stage j
B_j' = local backorders at stage j
B = customer backorders = B_J'
IT_j = inventory in transit to stage j (units in j's supply system)
I_j = echelon inventory at stage j = $I_j' + \Sigma_{i>j} (IT_i + I_i')$
IN_j = echelon net inventory at stage j = $I_j - B$.

Also, let

D_j =
leadtime demand for stage j, a generic random variable having the distribution of $D(L_j)$. The D_j are independent.

The cost factors are

b = backorder penalty-cost rate
h_j' = local inventory holding-cost rate at stage j
h_j = echelon inventory holding-cost rate at stage j = $h_j' - h_{j-1}'$,

where $h_0' = 0$.

The usual accounting scheme for in-transit inventories charges h_j' on IT_{j+1} as well as I_j'. We exclude such costs, in order to facilitate comparison among policies and systems. Thus, the total average cost, expressed in local terms, is

$$C(\mathbf{s}') = E[\Sigma_{j=1}^J h_j' I_j' + bB]. \tag{1}$$

The equivalent expression in echelon terms is

$$C(\mathbf{s}) = E[\Sigma_{j=1}^J h_j IN_j + (b + h_J')B] - E[\Sigma_{j=1}^J h_j' D_{j+1}]. \tag{2}$$

(Here, $D_{J+1} = 0$. The second term is necessary, because the first includes the usual in-transit holding cost, and $E[IT_j] = E[D_j]$.)

2.4. Local Policy Evaluation

For any policy $\mathbf{s}'$, the equilibrium local backorder variables satisfy the following recursion:

$$\begin{aligned} B_0' &= 0 \\ B_j' &= [B_{j-1}' + D_j - s_j']^+. \end{aligned} \tag{3}$$

And,

$$I_j' = s_j' - (B_{j-1}' + D_j) + B_j'. \tag{4}$$

(See, e.g., Graves 1985.) From these, we can compute $E[B]$ and $E[I_j']$ and thus the average cost [Equation (1)].

2.5. Echelon Policy Optimization

We now present a method to determine an optimal echelon base-stock policy, denoted $\mathbf{s}^*$. This is the Clark-Scarf algorithm, essentially as stated by Chen and Zheng:

Set $\underline{C}_{J+1}(x) = (b + h_J')[x]^-$. For $j = J, J - 1, \ldots, 1$, given $\underline{C}_{j+1}$, compute

$$\begin{aligned} \hat{C}_j(x) &= h_j x + \underline{C}_{j+1}(x) \\ C_j(y) &= E[\hat{C}_j(y - D_j)] \\ s_j^* &= \operatorname{argmin} \{C_j(y)\} \\ \underline{C}_j(x) &= C_j(\min\{s_j^*, x\}). \end{aligned} \tag{5}$$

At termination, set $C^* = C_1(s_1^*) - E[\Sigma_{j=1}^J h_j' D_{j+1}]$. This is the optimal cost.

A similar calculation can be used to evaluate any policy **s**. Just omit the optimization step, and use s_j in place of s_j^* in the last step. One can show that this method is equivalent to Equations (2) through (4). Conversely, one can show directly that Equation (5) optimizes over policies evaluated by Equations (2) through (4). This point underlies the extensions of §3. (To our knowledge, these observations are new here.)

Recursion (5) deserves to be called *the fundamental equation of supply-chain theory*. It captures the basic dynamics and economics of serial systems. It omits much, but any more comprehensive theory must build on it. We know little about its solution, however. The remainder of the paper begins to investigate it.

2.6. Decreasing Holding Costs

Examination of Equation (5) reveals that, for $j < J$, if $h_{j+1} \le 0$, then $s_{j+1}^* = \infty$, which implies $s_j'^* = 0$. In this case, we can eliminate stage j, replacing L_{j+1} by $L_{j+1} + L_j$ and h_{j+1} by $h_{j+1} + h_j$. (Rosling observes this.) Continue to eliminate stages in this way, until all the remaining $h_j > 0$. Thus, *a stage holds stock* ***only*** *when it is cheaper to hold it there than anywhere downstream.* This makes sense intuitively; downstream inventory provides more direct, effective protection against customer backorders than upstream inventory. The only possible advantage of upstream inventory is lower inventory-holding cost.

3. Other Demand and Supply Processes

The same methods can be used to evaluate and optimize, exactly or approximately, under a variety of other model assumptions.

3.1. Compound-Poisson Demand

Suppose that demand is a compound-Poisson process, and each increment of demand can be filled separately. *All* the results above remain valid. Here, each D_j has a compound-Poisson distribution, but that is the only difference.

3.2. Exogenous, Sequential Supply Systems

Consider a system like that of Svoronos and Zipkin (1991), specialized to a series structure: Each stage's supply system is stochastic. Stage j's system generates a *virtual leadtime* $L_j(t)$; a shipment to j initiated at time t arrives at $t + L_j(t)$. The system processes orders *sequentially*, so shipments arrive in the same sequence as the corresponding orders; that is, $t + L_j(t)$ is nondecreasing in t. Each supply system is *exogenous*, i.e., its internal state and $L_j(t)$ are stochastic processes, but they are unaffected by shipments. Each system is *ergodic*, i.e., $L_j(t)$ approaches a steady-state random variable L_j, regardless of initial conditions. Finally, these systems, and hence the $L_j(t)$ and L_j, are independent over j.

Svoronos and Zipkin show that Equations (3) through (4) evaluates a base-stock policy. Here, D_j has the distribution of $D(L_j)$, the demand over the (stochastic) virtual leadtime L_j, so $E[D_j] = \lambda E[L_j]$ and $V[D_j] = \lambda E[L_j] + \lambda^2 V[L_j]$. These $\mathrm{D_j}$ are again independent. Consequently, as explained in §2.5, Equation (5) finds the best base-stock policy.

3.3. Independent Leadtimes

Return to the Poisson-demand case. Suppose that each stage's leadtimes are i.i.d. random variables; in effect, each supply system consists of multiple identical processors in parallel. Let L_j be the generic leadtime random variable for stage j. In this context, Equations (3) through (4) remain valid with IT_j in place of D_j.

It is difficult to characterize the IT_j, in general. There is one case where it is easy, namely, when $\mathbf{s} = \mathbf{s}' = \mathbf{0}$. There, the system is equivalent to a tandem network of queues with Poisson input, where each node j has an infinite number of servers with service times L_j. So, IT_j has the Poisson distribution with mean $\lambda E[L_j]$, and the IT_j are independent. (See, e.g., Kelly 1979.). For general $\mathbf{s}' \ge \mathbf{0}$ we can use this same distribution to *approximate* the IT_j. This is, in fact, the key approximation underlying the METRIC procedure (see Sherbrooke 1968, 1986 and Graves 1985), specialized to series systems. It is quite accurate.

With this approximation, using D_j to stand for the approximate IT_j, Equations (3) through (4) evaluate a local policy. Therefore, Equation (5) computes the best base-stock policy, up to the approximation.

3.4. Limited-Capacity Supply Systems

Now, suppose each supply system consists of a single processor and its queue. The processing times at stage

j are i.i.d., distributed exponentially with rate μ_j. Assume $\lambda < \mu \equiv \min_j \{\mu_j\}$. Recursion (3), with IT_j in place of D_j, applies here too. Again, it is difficult to characterize the IT_j in general, but easy in the case $\mathbf{s}' = \mathbf{0}$. Here, IT_j has the geometric distribution with parameter $\rho_j = \lambda/\mu_j$, and the IT_j are independent (Kelly). This works well as an approximation for the general case, as shown by Buzacott et al. (1992), Lee and Zipkin (1992), and Zipkin (1995). So, Equation (5) again finds the (approximately) best base-stock policy.

4. Bounds and Heuristics

4.1. The Restriction-Decomposition Approximation

This section presents a fairly simple way to determine a useful heuristic policy and an upper bound on the optimal cost. The approach involves *restriction* of the policy space and *decomposition* of the resulting model into single-stage submodels. Accordingly, we call it the restriction-decomposition or RD approximation. This approach, or something like it, is widely used in practice. It is striking that this simple idea actually bounds the original system.

Let $\mathbf{J}_+$ be any subset of stages that includes J. We construct an approximation for any choice of $\mathbf{J}_+$ and then select the best $\mathbf{J}_+$. Index these stages in order by $j(m)$, $m = 1, \ldots, M$. So, $j(M) = J$. Also, denote $j(0) = 0$. Let

$$D_{(i,j]} = D_{i-1} + D_{i+2} + \ldots + D_j, \quad 0 \le i < j \le J,$$
$$D^m = D_{(j(m-1),j(m)]}, \quad m = 1, \ldots, M.$$

First, *restrict* $s_j' = 0, j \notin \mathbf{J}_+$, so that only stages in $\mathbf{J}_+$ are allowed to hold stock. Using Equation (3), one can readily show that

$$B'_{j(m)} = [B'_{j(m-1)} + D^m - s'_{j(m)}]^+, \ m = 1, \ldots, M.$$

The next steps effectively *decompose* the system at stages $\mathbf{J}_+$. It is easy to show that

$$B'_{j(m)} \le B'_{j(m-1)} + [D^m - s'_{j(m)}]^+, \quad m = 1, \ldots, M$$
$$B = B'_{j(M)} \le \textstyle\sum_{m=1}^{M} [D^m - s'_{j(m)}]^+$$
$$I'_{j(m)} \le [s'_{j(m)} - D^m]^+, \quad m = 1, , \ldots, M.$$

Consequently,

$$C(s') \le \textstyle\sum_{m=1}^{M} E[h'_{j(m)} \, [s'_{j(m)} - D^m]^+ + b[D^m - s'_{j(m)}]^+].$$

Equivalently, let

$$\hat{C}_j'(x) = h_j'[x]^+ + b[x]^-$$
$$C_{(i,j]}(y) = E[\hat{C}_j'(y - D_{(i,j]})], \quad i < j.$$

Then,

$$C(\mathbf{s}') \le \textstyle\sum_{m=1}^{M} C_{(j(m-1),j(m)]}(s'_{j(m)}).$$

Each term in this sum is the cost of a single-stage system. It charges the full penalty cost b to local backorders at each stage $j(m)$, while ignoring the effects of those backorders on downstream stages. In this sense it splits the system into separate subsystems.

Now, let $s_{(i,j]}$ minimize $C_{(i,j]}(y)$, and denote the minimal cost by $C^*_{(i,j]}$. Then,

$$C^* \le \textstyle\sum_{m=1}^{M} C^*_{(j(m-1),j(m)]}.$$

This relation holds for any $\mathbf{J}_+$. To find the best such bound over all possible $\mathbf{J}_+$, consider the following network: The nodes are $\{0,1, \ldots ,J\}$, the arcs are (i,j), $i < j$, and the arc lengths are $C^*_{(i,j]}$. The best bound, then, is the length of the shortest path from 0 to J. This problem has precisely the same structure as the dynamic economic lot-size problem of Wagner and Whitin (1958), and can be solved using the same algorithm.

From the best $\mathbf{J}_+$ one can construct a plausible heuristic policy: Set $s_j' = 0, j \notin \mathbf{J}_+$, and for $j = j(m) \in \mathbf{J}_+$, set $s_j' = s_{(j(m-1),j(m)]}$. The actual cost of this policy is no more than the upper bound. (Alternatively, use Equation (5) to find the optimal policy for the system restricted to $\mathbf{J}_+$. We have not tested this more refined approach.)

The RD approximation extends directly to the design problem: If there is a fixed cost k_j to build stage j, just add k_j to each $C^*_{(i,j]}$. Also, if several products share the network, compute the $C^*_{(i,j]}$ for each product, and then sum them over the products. The algorithm above then provides a heuristic solution and an upper bound.

We remark that the complexity of the RD heuristic appears to be $O(J^2)$, compared to $O(J)$ for the optimizing algorithm Equation (5). Indeed, we have observed that, for very large J, the heuristic can take longer than Equation (5). For smaller, plausibly-sized systems, however, the heuristic is usually much faster. And, it is a tractable method for the design problem.

Here is a further useful approximation: Scarf (1958) and Gallego and Moon (1993) show that

$$C^*_{(i,j]} \le (bh'_j)^{1/2}\sigma_{(i,j]} \equiv C^+_{(i,j]},$$

where $\sigma_{(i,j]}$ is the standard deviation of $D_{(i,j]}$. Using the $C^+_{(i,j]}$ in the calculations above yields a *distribution-free* bound, one that depends only on two moments of leadtimes and demands, not their actual distributions. Call this the *maximal RD approximation*. The same analysis yields a heuristic solution of the form $s^+_{(i,j]} = E[D_{(i,j]}] + z'_j\sigma_{(i,j]}$, where z'_j is a safety factor depending on b and h'_j, whose cost is no more than the upper bound. (This approach is much faster than the original RD heuristic, since $C^+_{(i,j]}$ is easier to compute than $C^*_{(i,j]}$.)

This simple form facilitates sensitivity analysis: Observe that, in the Poisson-demand, constant-leadtime case, each $C^+_{(i,j]}$ depends on λ through a factor $\sqrt{\lambda}$. Thus, the shortest path is independent of λ, and the cost bound is proportional to $\sqrt{\lambda}$. That is, *the heuristic's choice of stocking points is independent of the demand volume*, and *the true optimal cost is bounded above by a function proportional to* $\sqrt{\lambda}$. A similar analysis of $s^+_{(i,j]}$ suggests that the overall safety stock is proportional to $\sqrt{\lambda}$. The same is true for stochastic, independent leadtimes (§3.3). For exogenous, sequential leadtimes (§3.2), however, $\sigma_{(i,j]} = (\lambda E[L_{(i,j]}] + \lambda^2 V[L_{(i,j]}])^{1/2}$, so the optimal cost is bounded by a *linear* function of λ, as is the safety stock.

Likewise, the shortest path is independent of b, and the cost bound is proportional to $\sqrt{b}$.

The leadtime L_k affects $\sigma_{(i,j]}$ for all $(i,j]$ with $i < k \le j$. It has the biggest impact on the $\sigma_{(i,j]}$ for short intervals $(i,j]$ around k. Thus, for small k, L_k has a major impact only on terms $C^+_{(i,j]}$ with small j and hence low h'_j. Conversely, L_k for large k affects terms with large h'_j. This suggests that *downstream leadtimes have a greater impact on system performance than upstream ones.*

The familiar normal approximation yields an approximation to $C^*_{(i,j]}$ of the same form as $C^+_{(i,j]}$, namely, a factor depending on the cost parameters, times $\sigma_{(i,j]}$. It also yields a solution of the same form as $s^+_{(i,j]}$. Call this the *normal RD approximation*. So, the observations above about λ and the L_k remain valid. The cost factors, however, grow more slowly in b than $\sqrt{b}$.

Some additional bounds for two-stage systems can be found in Gallego and Zipkin (1994).

4.2. The Zero-Safety-Stock Heuristic

This approach (the ZS heuristic, for short) sets $s'_j = E[D_j]$, $j < J$, and then optimizes over s'_J. More precisely, to cover the case of non-integral $E[D_j]$, the heuristic sets $s'_j = \lceil \sum_{i \le j} E[D_i] \rceil - s'_{j-1}$, $j < J$. Then, using Equation (3), it computes the distribution of B'_{J-1}. Finally, it chooses s'_J to minimize the stage-J holding and penalty costs, a single-stage problem. (This method was inspired by some preliminary numerical results, in which the optimal s'_j was near $E[D_j]$, $j < J$.) Evidently, this is an $O(J)$ calculation, and it is very fast in practice.

4.3. The Two-Stage Heuristic

This approach (the TS heuristic) restricts inventory to two stages, the last one J and some single $j < J$. Given $j < J$, it finds the optimal policy for the resulting two-stage system. It then selects the best such policy over $j < J$. (This method too was based on empirical observations, namely, that restricting the number of locations sometimes has little cost impact.)

This technique requires solving $J - 1$ two-stage problems, nearly as much work as the full optimization algorithm Equation (5). The purpose of the heuristic is not speed. Rather, it is a tool to investigate stock-positioning issues: Where is stock most useful? And, how costly is the restriction to two stages? This approach also extends easily to the design problem; in that context it is a plausible heuristic for systems with large fixed facility costs.

5. Numerical Results

This section presents some numerical examples, to provide insight into the behavior of the optimal policy and the performance of the heuristics.

5.1. Specification

5.1.1. System Structure and Parameters. We assume Poisson demand and constant leadtimes. Without loss of generality, we fix the time scale so that the total leadtime is $L = 1$, and the monetary unit so that the last stage's holding cost is $h'_J = 1$. The stages are spaced symmetrically, so each stage j's leadtime is $L_j = 1/J$. We consider four numbers of stages, $J = 1, 4, 16, 64$; two demand rates, $\lambda = 16, 64$; and two penalty costs, $b = 9, 39$ (corresponding to fill rates of 90%, 97.5%).

5.1.2. Holding Cost Forms. We consider several forms of holding costs h'_j, depicted in Figure 1. The simplest form has *constant* holding costs, where all $h'_j = 1$. Here, there is no cost added from source to customer. This is a rather unrealistic scenario, but it is a useful starting point to help understand other forms. The *linear* holding-cost form has $h'_j = j/J$, or $h_j = 1/J$. Here, cost is incurred at a constant rate as the product moves from source to customer. This is quite realistic. *Affine* holding costs, where $h'_j = \alpha + (1 - \alpha)j/J$ for some $\alpha \in (0,1)$, are even more realistic. Here, the material at the source has some positive cost, and the system then adds cost at a constant rate. This form is a combination of the constant and linear forms. In Figure 1 and the calculations below, $\alpha = 0.75$.

The last two forms represent deviations from linearity. The *kink* form is piecewise linear with two pieces. The system incurs cost at a constant rate for a while, but at some point shifts to a different rate, which remains constant from then on. Here, the kink occurs halfway through the process, at stage $J/2$. So, for some $\alpha \in (-1,1)$, $h_j = (1 - \alpha)/J$, $j \leq J/2$, and $h_j = (1 + \alpha)/J$, $j > J/2$. Again, we set $\alpha = 0.75$. Finally, in the *jump* form, cost is incurred at a constant rate, except for one stage with a large cost. Here, the jump occurs just after stage $J/2$. So, $h_j = \alpha + (1 - \alpha)/J$, $j = J/2 + 1$, and $h_j = (1 - \alpha)/J$ otherwise, for some $\alpha \in (0,1)$. We can view this as linear cost before $J/2$ and affine cost after. Here again, $\alpha = 0.75$.

5.2. Optimal Policy

5.2.1. Constant Holding Costs. The optimal policy in this case is simple: For $j < J$, $s_j'^* = 0$; only the last stage carries inventory. Stage J, in effect, becomes a single-stage system with leadtime L. The optimal policy is the same for all J. This is also the optimal policy for $J = 1$ under any other holding-cost form.

5.2.2. Linear Holding Costs. Figure 2 shows the optimal policy $\mathbf{s}^*$ for $J = 64$ and two values each of λ and b. Several observations are worth noting: The curves are *smooth* and *nearly linear;* the optimal policy does *not* lump inventory in a few stages, but rather spreads it quite evenly. The departures from linearity are interesting too: The curves are *concave.* Thus, the policy focuses safety stock at stages nearest the customer.

5.2.3. Affine Holding Costs. Figure 3 shows the optimal policy. For $j > 1$, the curves follow the same pattern as in Figure 2. (Indeed, the curves for $b = 9$ here are identical to those for linear costs and $b = 39$, because these two cases have identical ratios $h_j/(b + h'_j)$, $j > 1$.) However, the curves break down sharply at $j = 1$ (because h_1 is large). Therefore, the equivalent policy $\mathbf{s}^-$ is flat for small j, and so the policy holds *no* inventory at early stages. This solution is intermediate between those for constant and linear costs. As α increases and the costs move upwards, stocks shift toward the customer. The total system stock decreases

Figure 1 Holding Cost Forms

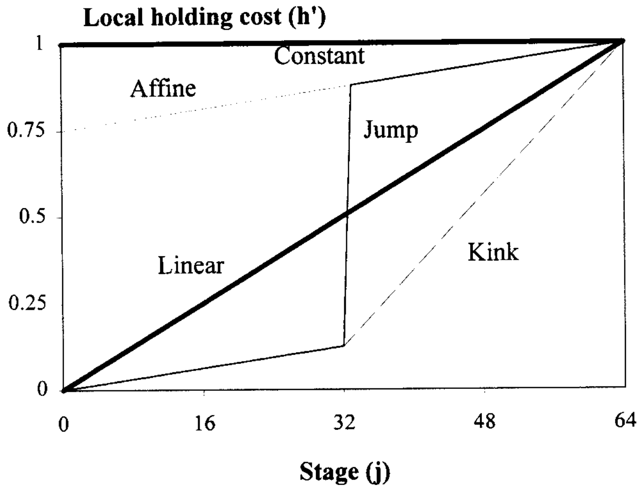
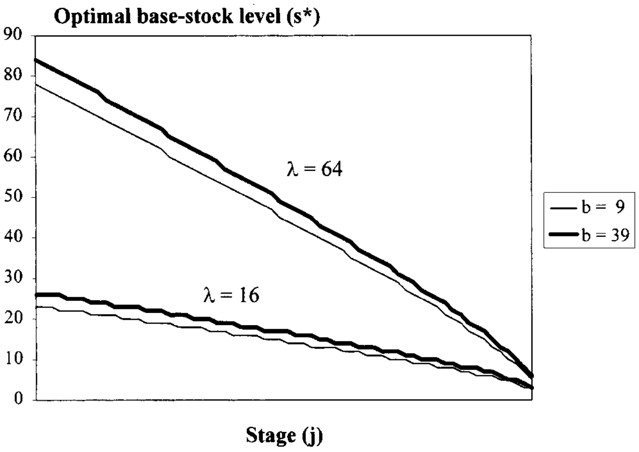


Figure 2 Optimal Policy: Linear Holding Costs



slightly. But, perhaps surprisingly, stocks near the customer actually increase.

5.2.4. Kink Holding Costs. Figure 4 displays **s***. Downstream from the kink (before Algorithm (5) encounters it), the curves exhibit the same pattern as in the linear case. Upstream from the kink, the policy again follows the linear pattern, almost as if the kink were the last stage. The net result is substantial stock at and just before the kink, where holding costs are low relative to later stages.

5.2.5. Jump Holding Costs. Figure 5 displays **s***. From the jump on, the policy behaves much as in the affine case: smooth, concave decrease beyond the jump, but a sharp break downwards at the jump. Upstream from the jump, the policy again follows the pattern of the linear case. Thus, there is substantial stock just before the jump and none just after it.

5.3. Sensitivity Analysis

5.3.1. Number of Stages. Figure 6 compares the **s*** for different *J*s, each with linear holding costs, $\lambda = 64$, and $b = 39$. The curves follow the same patterns as before, as closely as the restricted numbers of stages allow. Indeed, the actual echelon stock at a stocking point is nearly identical to the $J = 64$ case. Closer inspection shows that the total system stock is slightly

Figure 3 Optimal Policy: Affine Holding Costs

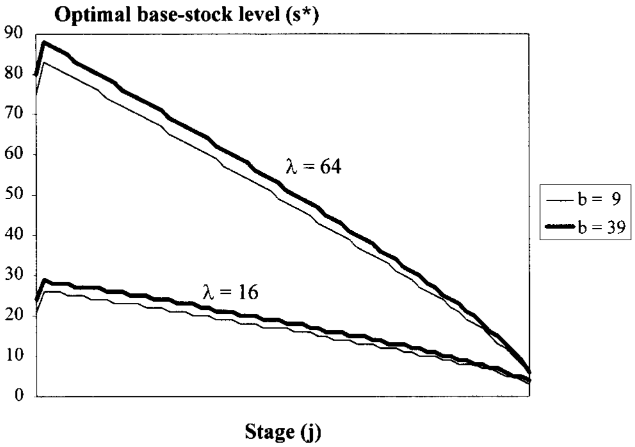


Figure 4 Optimal Policy: Kink Holding Costs

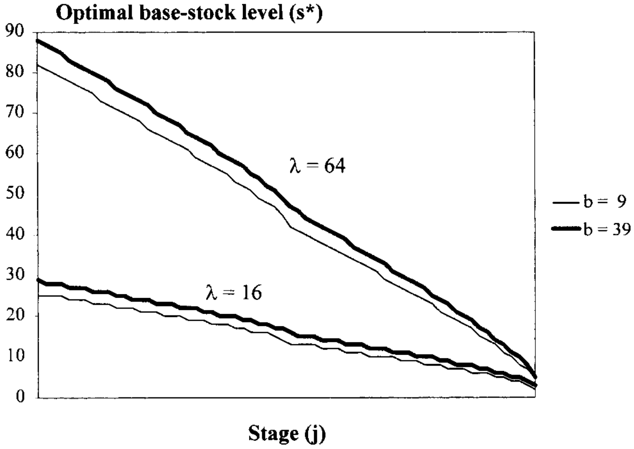


Figure 5 Optimal Policy: Jump Holding Costs

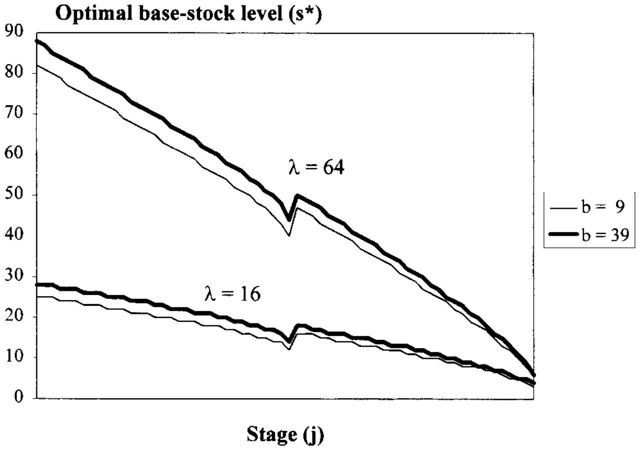
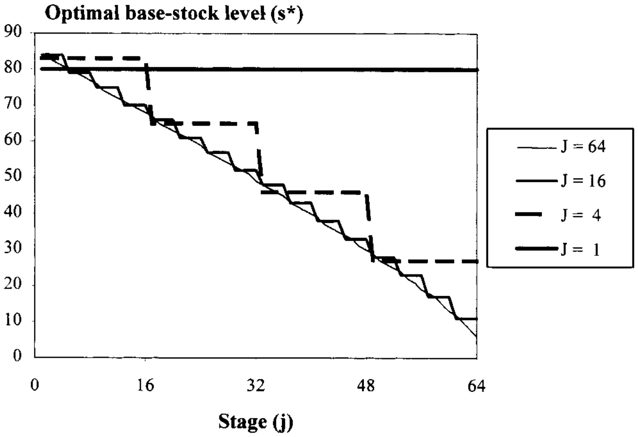


Figure 6 Optimal Policy: Effects of *J*



higher for larger J. Likewise, the optimal cost decreases in J, but quite slowly, as shown in Figure 7.

Similar results hold for affine holding costs. Indeed, the optimal cost is even less sensitive to J. For kink holding costs (Figure 8), the optimal cost is significantly lower at $J = 4$ than at $J = 1$, due to the availability of the low-cost stocking point at the kink. Larger Js yield relatively minor improvements. The jump form displays a similar pattern. Thus, for these two forms, it is important to position stock at the kink (or jump). Otherwise, the cost is quite insensitive to J.

These results suggest that *the system cost is relatively insensitive to stock positioning*, provided the overall stock level is about right, and obvious low-cost stocking points are exploited. We shall see further evidence for this below.

Figure 7 Optimal Cost: Linear Holding Costs

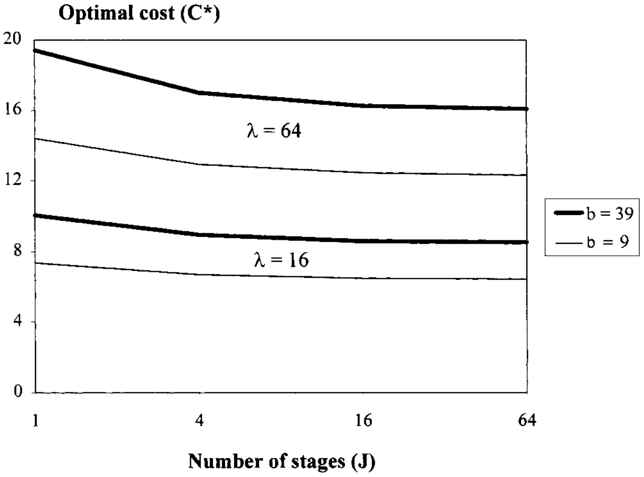
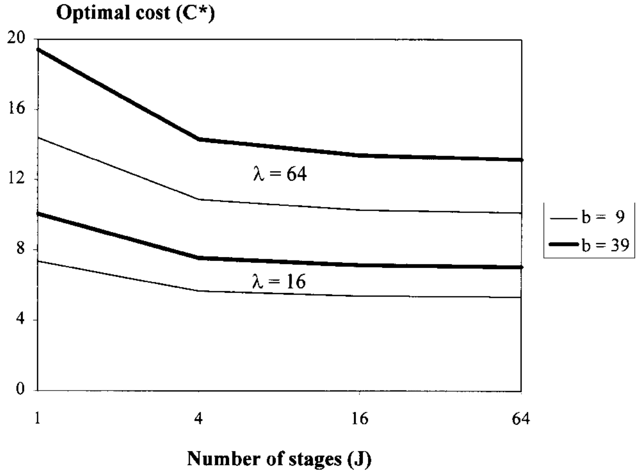


Figure 8 Optimal Cost: Kink Holding Costs



5.3.2. Demand Rate.

In Figures 7 and 8 the optimal cost for $\lambda = 64$ is about twice that for $\lambda = 16$ in every case. This is consistent with the notion that the optimal cost is nearly proportional to $\sqrt{\lambda}$, as suggested in §4.1. We have also plotted, but omit here, the cumulative safety stocks $\Sigma_{i \le j}\, s_j'^{*} - \lambda j/J$. The curves for $\lambda = 64$ are about twice those for $\lambda = 16$. So, the safety stocks too are nearly proportional to $\sqrt{\lambda}$.

5.3.3. Backorder Cost.

The figures above indicate that the base-stock levels and optimal cost are increasing in b. The policy, however, is not very sensitive to b. The cost, though rather more sensitive, grows considerably slower than $\sqrt{b}$, as suggested by the normal RD approximation.

5.3.4. Leadtimes.

Figures 7 and 8 provide some evidence for the notion that downstream leadtimes are more important than upstream ones. Starting with linear holding costs, contract the downstream leadtimes and expand the upstream ones, keeping L and the h_j' fixed. The result looks much like the kink form with $\alpha \in (0,1)$. And, the kink form has lower optimal cost for $J > 1$.

5.4. Performance of Bounds and Heuristics

5.4.1. The RD Approximation.

Figure 9 shows the policies chosen by the RD heuristic in one case ($J = 64$, $\lambda = 64$, $b = 39$) for all four holding-cost forms. (The same policy is chosen for the kink and jump forms.) These policies are quite different from the corresponding optimal ones; they concentrate stock in just a few stages. For the linear form, the policy places a small inventory near the source (9 units at stage 3) and a large one (77) at the last stage. For the affine form, the policy is even more extreme, placing all its stock (80) at the end. For the kink and jump forms, the policy places substantial inventory (46) at stage 32, just before the cost increase, a little near the source (9 at stage 2), and the rest (44) at the end. Also, the total system stocks are slightly larger than optimal. The results for other J, λ, and b are similar.

Figure 9 RD Heuristic's Policies

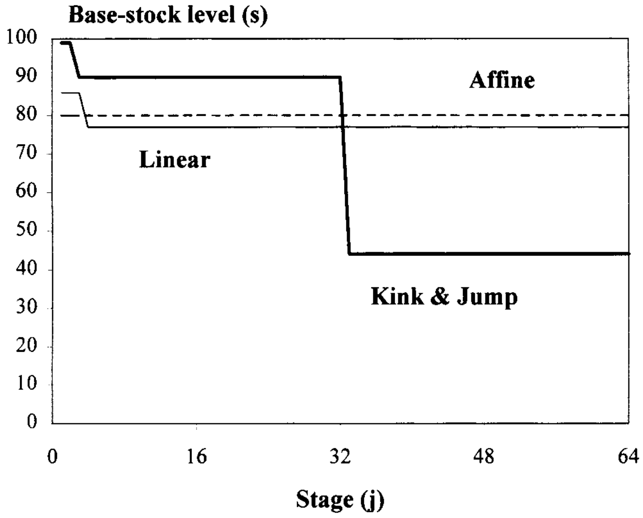


Table 1 Heuristics' Percentage Costs over Optimal

Form	RD	ZS	TS
Linear	10–20%	2–8%	4–11%
Affine	1–3%	3–14%	0–2%
Kink	9–22%	11–25%	5–17%
Jump	5–7%	11–15%	1–3%

Even so, the RD heuristic and the cost bound perform fairly well. Table 1 shows the percentage errors for all three heuristics. For example, for the linear form, the RD policy's cost exceeds the optimal by 10%–20%. (The errors tend to increase slowly in J, λ, and b.) These errors are far smaller than the cost differences between systems. The cost bound is usually just a bit more than the actual heuristic policy's cost.

Thus, the RD approximation provides crude but robust estimates of system performance. It is certainly accurate enough for rough-cut design studies. This fact, coupled with the gross differences between the RD and optimal policies, is further evidence of the insensitivity of performance to stock positioning.

5.4.2. The ZS Heuristic. The ZS heuristic, by definition, sets s_j' to the average leadtime demand up to the last stage. It sets s_J' larger than the optimal policy does, to compensate for the lower stocks at earlier stages. It generates the same policy for all four cost forms. It works very well for linear holding costs, rather less well for affine costs, and not so well for the kink and jump forms.

5.4.3. The TS Heuristic. For $J = 64$, for linear costs, the TS heuristic places stock just past the middle of the system, in addition to stage J. Specifically, for $\lambda = 64$ and $b = 39$, it chooses $j = 36$. For affine costs, the heuristic places stock further downstream, at $j = 48$. (The locations are just slightly different for the other λ and b.) For the kink and jump forms, it selects $j = 32$, just before the cost increase, in all cases. The results are similar for smaller J. As Table 1 indicates, this method performs quite well; it is the best among the three heuristics. This is yet more evidence of the insensitivity of performance to stock positioning.

6. Conclusions

We have seen that the optimal policy depends on the growth of holding costs between source and customer. For constant costs, the policy puts all stock at the last stage. For linear costs, the policy distributes stock quite evenly, though favoring downstream sites. In other cases the policy can be understood as a systematic combination and variation of these patterns.

On the other hand, although it is important to optimize the system-wide inventory and to exploit especially low holding costs, system performance is otherwise fairly insensitive to stock positioning. One can deviate substantially from the optimal policy for a rather small cost penalty, as in the restriction to smaller J and the heuristics. In particular, the RD and TS heuristics work fairly well; they capture the gross behavior of the optimal policy, though differing substantially in detail. Consequently, they are reasonable heuristics for the design problem.

The sensitivity of the system to its parameters is similar in many ways to the familiar single-stage system. For instance, with constant leadtimes, the optimal cost and safety stocks increase as the square root of the demand rate. Multistage systems have certain additional characteristics, however. For example, downstream leadtimes have greater impacts on performance than upstream ones.

We have presented these results to several groups of managers in different industries. Their reactions are worth reporting. They showed considerable interest in the *forms* of the figures as diagnostic devices. For example, they wanted to plot their own holding costs in the style of Figure 1, to see where cost accrues quickly and where slowly. (This type of diagram is called a *time-cost profile* by Fooks (1993) and Schraner (1994). Observe that the in-transit holding cost is essentially the area under each curve.) Likewise, a plot of actual inventories in the manner of Figures 2 through 5 is a convenient way to see just where stock is concentrated.

Many managers at first resisted the notion that stock should be concentrated close to customers. After all, the downstream sites are the most expensive ones. But, following discussion of the sites' different degrees of stockout protection, as in §2.6, most agreed that the optimal policy was at least plausible. Several noted that their own firms' stock-positioning policies were quite different, and planned to investigate the alternative suggested by the model. Similarly, many had embraced the idea of reducing total leadtime, and were dubious that downstream leadtimes could be more important than upstream ones. Once the logic was explained, however, they accepted it.

Finally, none of the managers found it hard to believe that the heuristics perform well. Indeed, they preferred solutions that concentrate stock in only a few locations, and they appreciated the simplicity of the heuristics. Their experience suggested that *all* real systems incur some fixed costs, as in the design problem.

Several questions remain: Are there better heuristics? Do the results extend to more complex systems, such as distribution systems and systems with fixed order costs? These are subjects of ongoing research.[1]

References

Arrow, K., S. Karlin, H. Scarf, eds. 1958. *Studies in the Mathematical Theory of Inventory and Production.* Stanford University, Stanford, CA.

Axsäter, S. 1993. Continuous review policies for multi-level inventory systems with stochastic demand. S. Graves, A. Rinnooy Kan, P. Zipkin, eds. *Logistics of Production and Inventory.* Elsevier (North-Holland), Amsterdam, The Netherlands. Chapter 4.

——, K. Rosling. 1993. Installation vs. echelon stock policies for multilevel inventory control. *Management Sci.* **39** 1274–1280.

Buzacott, J., S. Price, J. Shanthikumar. 1992. Service level in multistage MRP and base stock controlled production systems. G. Fandel, T. Gulledge, A. Jones, eds. *New Directions for Operations Research in Manufacturing.* Springer, Berlin, Germany.

Chen, F., Y. Zheng. 1994. Lower bounds for multi-echelon stochastic inventory systems. *Management Sci.* **40** 1426–1443.

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

Federgruen, A. 1993. Centralized planning models for multi-echelon inventory systems under uncertainty. S. Graves, A. Rinnooy Kan, P. Zipkin, eds. *Logistics of Production and Inventory.* Elsevier (North-Holland), Amsterdam, The Netherlands. Chapter 3.

——, P. Zipkin. 1984. Computational issues in an infinite-horizon, multiechelon inventory model. *Oper. Res.* **32** 818–836.

Fooks, J. 1993. *Profiles for Performance.* Addison-Wesley, New York.

Gallego, G., I. Moon. 1993. The distribution-free newsboy problem: Review and extensions. *J. Oper. Res. Soc.* **44** 825–834.

——, P. Zipkin. 1994. Qualitative analysis of multi-stage production-transportation systems: Stock positioning and performance estimation. Working paper, Columbia University, New York.

Graves, S. 1985. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Sci.* **31** 1247–1256.

——, A. Rinnooy Kan, P. Zipkin, eds. 1993. *Logistics of Production and Inventory.* Handbooks in Operations Research and Management Science, Volume 4, Elsevier (North-Holland), Amsterdam, The Netherlands.

Gross, D., R. Soland, C. Pinkus. 1981. Designing a multi-product, multi-echelon inventory system. L. Schwarz ed. *Multi-Level Production/Inventory Control Systems: Theory and Practice.* North-Holland, Amsterdam, The Netherlands. Chapter 1.

Kelly, F. 1979. *Reversibility and Stochastic Networks.* Wiley, New York.

Langenhoff, L., W. Zijm. 1990. An analytical theory of multi-echelon production/distribution systems. *Statist. Neerlandica* **44** 3, 149–174.

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

Nahmias, S. 1981. Managing reparable item inventory systems: A review. L. Schwarz ed. *Multi-Level Production/Inventory Control Systems: Theory and Practice.* North-Holland, Amsterdam, The Netherlands. Chapter 13.

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

Scarf, H. 1958. A min-max solution of an inventory problem. K. Arrow, S. Karlin, H. Scarf, eds. *Studies in the Mathematical Theory of Inventory and Production.* Stanford University, Stanford, CA. Chapter 12.

Schraner, E. 1994. Optimal production operations sequencing. Working paper, Stanford University, Stanford, CA.

Schwarz, L., ed. 1981. *Multi-Level Production/Inventory Control Systems: Theory and Practice.* North-Holland, Amsterdam, The Netherlands.

Sherbrooke, C. 1968. METRIC: A multi-echelon technique for recoverable item control. *Oper. Res.* **16** 122–141.

[1] We are grateful to Jing-Sheng Song for helpful comments on earlier versions of this paper.

——. 1986. VARI-METRIC: Improved approximations for multi-indenture, multi-echelon availability models. *Oper. Res.* **34** 311–319.

Svoronos, A., P. Zipkin. 1991. Evaluation of one-for-one replenishment policies for multiechelon inventory systems. *Management Sci.* **37** 68–83.

Wagner, H., T. Whitin. 1958. Dynamic version of the economic lot size model. *Management Sci.* **5** 89–96.

Zipkin, P. 1995. Processing networks with planned inventories: Tandem queues with feedback. *European J. Oper. Res.* **80** 344–349.

Accepted by Stephen Graves; received November 18, 1996. This paper has been with the authors 8 $\frac{1}{2}$ months for 4 revisions.