Manufacturing & Service Operations Management

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

Performance Evaluation and Stock Allocation in Capacitated Serial Supply Systems

Diwakar Gupta, N. Selvaraju,

To cite this article:
Diwakar Gupta, N. Selvaraju, (2006) Performance Evaluation and Stock Allocation in Capacitated Serial Supply Systems. Manufacturing & Service Operations Management 8(2):169-191. http://dx.doi.org/10.1287/msom.1060.0104

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 8, No. 2, Spring 2006, pp. 169–191
ISSN 1523-4614 | EISSN 1526-5498 | 06 | 0802 | 0169

informs

DOI 10.1287/msom.1060.0104

Performance Evaluation and Stock Allocation in Capacitated Serial Supply Systems

Diwakar Gupta
Industrial and Systems Engineering, Department of Mechanical Engineering, University of Minnesota, 111 Church Street, SE, Minneapolis, Minnesota 55455, guptad@me.umn.edu

N. Selvaraju
Department of Mathematics, Indian Institute of Technology, North Guwahati, Guwahati Assam, 781034, India, nselvaraju@iitg.ernet.in

We develop an approximation scheme for performance evaluation of serial supply systems when each stage operates like a single-server queue, and its planned inventories are managed according to a base-stock policy. We also present a near-exact matrix-geometric procedure for benchmarking our approximation relative to two other methods proposed in the literature. Through numerical tests, we demonstrate that our method is superior, both for performance estimation and for policy parameter optimization. Using this technique, we then perform experiments that address the following issues. What proportion of the optimal total inventory should managers allocate to upstream production stages to minimize the sum of inventory and backorder costs? If managerial action could lower holding cost rate or add capacity, which stages of the supply system should be targeted for maximum net benefit? Such concerns have been the subject of several recent studies relating to supply networks with constant and random independent lead times. We shine light on optimal actions for serial supply systems that experience congestion.

Key words: capacitated supply systems; queues with planned inventories; approximations; service constraints; matrix-geometric procedure
History: Received: July 25, 2002; accepted: September 3, 2005. This paper was with the authors 19 months for 3 revisions.

1. Introduction

This article is concerned with tandem supply networks with planned internal inventories, the likes of which arise naturally in instances ranging from process industries to discrete parts manufacturing. Each node or stage in this network represents an activity comprising one or more of manufacturing, assembly, warehousing, and transportation operations. Activities are modeled as single-server queues. This reflects the fact that processing capacity is limited, congestion effects are significant, and orders are processed sequentially. In these systems, only the farthest downstream stage experiences external customer orders. However, stock can be maintained at any stage. More stock in downstream stages generally means lower customer lead times, but greater holding costs.

Local inventories in the supply networks are often controlled by a base-stock control policy. According to this policy, each stage has an inventory buffer, whose size is called the local base-stock level. Items stored in the buffer have completed processing at that stage. Each stage continues to produce, provided materials are available at the immediate upstream stage, until its buffer is full. Each stage maintains a log of unfilled requests from downstream stages; backorders are filled on a first-come-first-served (FCFS) basis whenever an item completes processing. Backorders at the farthest downstream stage equal the number of customers that are delayed for lack of inventory. Thus, inventory holding costs may be incurred at each stage, but customer backorder costs are incurred only at the last stage (see §2 for further details). We are interested in determining system performance measures such as average inventory and backorder levels for specified base-stock quantities, as well as the optimal base-stock levels. The latter minimize the sum of

supply-chain-wide inventory holding and customer backorder costs.

Although a base-stock control policy is not necessarily optimal for the limited capacity systems of interest to us, it is often used to control series supply systems because of the ease of implementation and the fact that it performs well relative to other heuristic policies. Furthermore, an echelon base-stock policy is known to be optimal when the supply lead time at each stage either is constant or generated by an exogenous sequential supply system (see Zipkin 2000, §8.3), both of which are related to the system we study in this paper. (Echelon base-stock level is the sum of local base-stock level at that stage and all downstream stages. It is well known that for the systems under investigation in this article there is an equivalent local base-stock policy for each echelon base-stock policy, and vice versa.)

Interest in modeling serial supply systems dates back to the seminal paper of Clark and Scarf (1960). They prove that the optimal policy for series processing networks with linear costs and constant supply lead times is an echelon base-stock control policy. Further refinements are carried out by Federgruen and Zipkin (1984), Rosling (1989), and Chen and Zheng (1994). Fill-rate-based performance analysis and policy optimization of base-stock controlled series supply chains has also been studied (see, for example, Boyaci and Gallego 2001, Sobel 2004, and references therein). Gallego and Zipkin (1999) and Zipkin (2000, §8.3) provide an excellent survey of relevant literature.

Whereas a majority of the above-mentioned studies pertain to supply systems that have either constant or random independent supply lead times, Lee and Zipkin (1992) (hereafter referred to as LZ) and Buzacott et al. (1992) (hereafter referred to as BPS) provide approximate methods to evaluate the performance of capacitated supply systems with specified base-stock levels. In this situation, supply lead times are both random and correlated across stages due to congestion effects. Consequently, an exact computation of the supply system's performance is difficult.

The primary objective of this article is to develop an improved performance evaluation technique for these types of supply systems, which leads to better decisions. We test our approach by comparing it to a matrix-geometric-analysis based near-exact numerical procedure for estimating mean inventory and backorder levels. Comparisons reveal that our approximation reduces the average estimation error by about half in a wide range of numerical experiments as compared with the approximation that treats each stage as an independent $M/M/1$ queue (called the BPS-LZ approximation). We prove that the BPS-LZ approximation actually provides an upper bound on a key measure of congestion in two-stage supply chains. Numerical experiments, furthermore, confirm that the BPS-LZ approximation continues to overestimate congestion levels in series systems with multiple stages. We also compare our approach to a methodology based on a suggestion by BPS for a two-stage system (called the BPS approximation) that treats Stage 2 and all downstream stages as $GI/M/1$ queues. Numerical experiments show that the BPS approximation underestimates the key performance measure when stocking levels are small, whereas our technique corrects this problem to a large extent. We use our technique in numerical experiments to optimize base-stock levels under different parameter values, and show that it finds the true optimal stocking levels more often than the BPS-LZ or BPS approximations do. Moreover, when it fails to identify the true optimal solution, its solution is generally less costly than the other two methods.

A second objective of this study is to develop a deeper qualitative understanding of the forces that affect stock positioning in serial supply systems. Gallego and Zipkin (1999) address precisely these issues when lead times are deterministic. Through numerical examples in which lead time at a stage equals $1/J$, in a J-stage system, they show the following important results (see also Zipkin 2000, §8.3):

1. By comparing optimal costs in systems with J stages (stocking points), for different values of J, Gallego and Zipkin (1999) observe that the system cost is relatively insensitive to the value of J. Because a larger value of J corresponds to more choices for placing inventory, they conclude that as long as obvious low-cost stocking points are exploited and the overall stock level is about right in each instance, the system cost is relatively insensitive to stock positioning.

2. In a two-stage system, Stage 1 should hold inventory either when its inventory holding costs are lower, or when its lead time is substantial relative to Stage 2's lead time. In all other cases, inventory should be pushed downstream to Stage 2. Similar observations also extend to systems with more than two stages.

3. More safety stock should be placed at downstream stages, i.e., closer to the customers.

Gallego and Zipkin (1999) state that virtually all their results remain valid for a variety of more complex supply systems with stochastic lead times, including supply systems with limited capacity.

We test whether the above-mentioned guidelines hold when the supply chain is capacity constrained. However, our approach to the question of stock allocation is slightly different from Gallego and Zipkin's approach. In each comparison, we keep the number of stages fixed and ask how sensitive the system cost is to the relative amount of inventory held at each stage when the total inventory in the supply chain is fixed at the optimal level. We are motivated to pose the question in this manner for two reasons. First, many serial supply systems have a common storage space for both semifinished and finished goods inventories. Companies use markers (e.g., painting space for different products in different colors) to allocate this space to different semifinished and finished goods. It is therefore managerially relevant to ask what proportion of total storage space should be allocated to semifinished inventory at each stage of processing, given that the total amount of space available is optimal. We believe less-accurate performance evaluation algorithms can identify approximately the right amount of total inventory, but that its allocation may vary sharply depending on the accuracy of the algorithm. Therefore, the second reason for choosing this method of framing the question is to highlight a key difference between our approximation and the approximations proposed earlier. Our method tends to position more inventory upstream. By normalizing with respect to the optimal inventory level, we are able to isolate the effect of the performance evaluation method on inventory allocation from its effect on the total amount of inventory.

Shang and Song (2003a) develop bounds and a heuristic procedure for finding the optimal echelon base-stock levels for series supply chains with constant lead times. They study the effect of system parameters, such as the location of bottlenecks and the relative magnitude of holding costs, on the choice of base-stock levels. Shang and Song also evaluate the relative benefit of reducing echelon holding cost rate and supply lead time at each stage. Reduction in holding cost is possible by introducing new technology, applying better management, or outsourcing. They show that reducing echelon holding cost rate at the farthest upstream stage and reducing supply lead time at the farthest downstream stage are optimal.

We confirm that many findings in Gallego and Zipkin (1999) and Shang and Song (2003a) also hold for capacitated series systems. However, there are some differences. For example, in our experiments, Stage 1 holds inventory even if its utilization (and hence supply lead time) is not higher than Stage 2's, and Stage 1's holding cost rate is only slightly smaller than Stage 2's. In some examples, Stage 1 holds inventory when holding cost rates are identical and Stage 1's utilization is only slightly larger than Stage 2's. This seems to contradict Gallego and Zipkin, who recommend placing all inventory downstream under similar circumstances. However, the overall qualitative behavior of the multistage systems agrees with the conclusions reached in Gallego and Zipkin; that is, more stock, including safety stock, should be held closer to the customer unless there are stocking points farther upstream with a substantially lower holding cost rate, or an upstream stage is severely capacity constrained.

When we vary the amount of stock at each stage in a two-stage capacitated system, while keeping the sum of the two stock levels equal to the optimal total amount of stock, we find that the system cost is insensitive to stock positioning only in a certain neighborhood of the optimal stock at each stage. (Note that this observation is not inconsistent with Gallego and Zipkin's 1999 conclusion regarding stock positioning. Because they assume optimal distribution of stock to different stages, and then vary the number of stages, the cost of suboptimal allocation is not considered in their comparisons.) In fact, the cost is sharply higher when Stage 1 carries more than the optimal amount of stock at the expense of smaller-than-optimal amount

of stock at Stage 2. Essentially, when holding inventory at Stage 1 is cheaper, optimal stocking level at Stage 1 is already high. In this situation, shifting a small amount of additional inventory from Stage 2 to Stage 1 can substantially increase backorder costs, whereas the concomitant holding cost savings are small. Note that if additional inventory is added to Stage 1 without a simultaneous decrease in Stage 2 inventory, i.e., total inventory is increased, then the impact of this action on system cost may be small under similar parameter values.

We also carried out experiments similar to Shang and Song (2003a) to study the effect of adding capacity. The results reported in this paragraph pertain to balanced systems (equal stage utilizations). We also studied systems with unequal utilizations, but in those cases the patterns of behavior are not as clear. When processing capacity can be increased by an equal amount at one or more stages that have equal processing capability to begin with, we find that if holding costs are either constant, or a linear or affine function of stage index, capacity should be applied to create two subsystems. Processing stages within each subsystem have equal utilization, and the later occurring subsystem has smaller utilization. This observation does not always agree with Shang and Song (2003a), who find that in uncapacitated systems, lead time reduction should be applied to the farthest downstream stage. Note that increasing processing capacity has the effect of stochastically reducing lead time at the affected stage in our model.

An alternative formulation of the problem under investigation in this article minimizes average inventory costs under a service-level constraint. These service-constrained models are of interest to managers who find it easier to specify service goals in terms of measures such as fill rate rather than to specify backorder costs (*fill rate* is the proportion of demands met immediately from on-hand inventory). Boyaci and Gallego (2001) provide a detailed account of literature on service-constrained optimization models for multistage systems, and Shang and Song (2003b) and Sobel (2004) provide examples of some recent work. We use our approximation scheme and carry out numerical experiments for a fill-rate-constrained two-stage system with limited capacity. The BPS-LZ approximation underestimates the true fill rate, whereas the estimates under our approximation are closer to the true values as compared with the BPS-LZ and BPS approximations. Our approximation also performs better in predicting the true optimal base-stock levels as compared with the existing approximations.

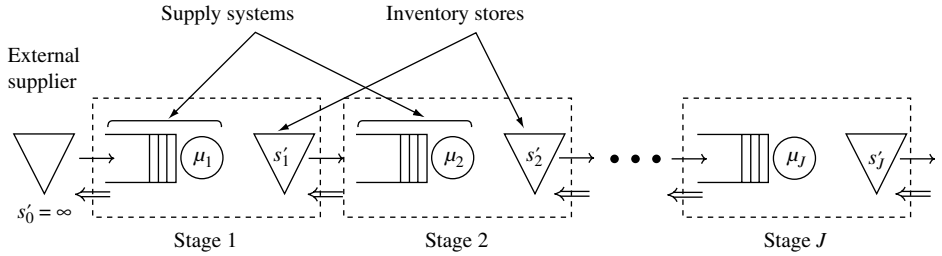
The remainder of this article is organized as follows. In §2 we introduce notation, specify a mathematical model, and outline a methodology based on matrix-geometric arguments for numerically computing the joint distribution of queue lengths at each stage of the supply chain. Our approximation procedure and its comparison to the BPS-LZ and BPS methods are reported in §3. Section 4 considers optimization of the policy parameters (base-stock levels). Sections 4 and 5 provide results of experiments that parallel those reported for systems with constant lead times in Gallego and Zipkin (1999) and Shang and Song (2003a). The new approximation scheme is extended to a fill-rate-constrained model in §6 for performance evaluation and policy optimization. Finally, we present concluding remarks in §7. Proofs of our propositions are in the appendix.

2. Preliminaries: Notation, Model, and Queue Occupancy Levels

We consider a network of J stages or nodes. Each stage has a supply system (processing facility) that is modeled as a queue of infinite capacity and a single server with an exponential processing time of mean $1/\mu_j$, $j = 1, 2, \ldots, J$. Stage inventories are controlled by a base-stock policy, specified by the nonnegative integer parameters $s'_j \geq 0$, $j = 1, 2, \ldots, J$, representing the local base-stock levels. Units move from an inventory store to the next stage in response to demands. Customer demands occur one unit at a time according to a Poisson process with rate λ. stage J fills customer demand if there is a finished unit available; otherwise, the demand is backordered. A demand at stage j always places an order at the predecessor stage $(j-1)$. This means that a customer demand creates an equivalent demand at every stage. The external supplier, represented by Stage 0, is assumed to possess an unlimited store of raw material kits. Put differently, $s'_0 = \infty$, and the external supplier releases a unit of raw material to Stage 1 as soon as the demand occurs.

If there are backorders at a stage when that stage completes processing of an item, then the completed

Figure 1 Schematic of the Series Supply Network



Note. Forward arrows (⟶) indicate flow of materials and backward arrows (⟸) indicate propagation of demand information.

unit is transferred immediately to fill a pending order in the FCFS manner. Thus, each stage tries to work down its backorder level first and then fill its inventory store up to s'_j. The supply lead time at stage j is defined as the time elapsed from the moment a unit is released by stage $(j-1)$ to the instant that particular unit reaches stage j inventory store. Propagation of demand information is assumed to occur instantaneously. A schematic of this system is shown in Figure 1.

Let h'_j denote the local inventory holding cost rate at stage j, and let b denote the per-unit backorder penalty rate (for customer demands). The echelon inventory holding cost rate at stage j is $h_j = h'_j - h'_{j-1}$ (with $h'_0 = 0$). The use of prime notation to denote local parameters is consistent with the standard usage in literature (see, for example, Zipkin 2000). Echelon base-stock policy is denoted by parameters s_j, $j = 1, \ldots, J$, where for a given sequence $(s'_j)_{j=1}^J$, $s_j = \sum_{i \geq j} s'_i$. We define the following random variables representing equilibrium distributions of the corresponding quantities:

N_j = Number of units in stage j's supply system (with $N_{J+1} \equiv 0$)
B_j = Number of units backordered at stage j
I'_j = Number of units in the inventory store at stage j
K_j = Number of units needed to be produced at stage j to bring its inventory level to s'_j. $K_j = s'_j - I'_j + B_j$
B = Customer backorders $(=B_J)$
I_j = Echelon inventory at stage $j = I'_j + \sum_{i>j}(N_i + I'_i)$
IN_j = Echelon net inventory at stage $j = I_j - B_j$

Using the above notation, the total average cost is

$$C(\mathbf{s}') = E\left[\sum_{j=1}^{J} h'_j(I'_j + N_{j+1}) + bB\right]$$

$$= E\left[\sum_{j=1}^{J} h_j \mathrm{IN}_j + (b + h'_J)B\right], \qquad (1)$$

where the last equality is obtained after recasting the cost expression in echelon terms (Zipkin 2000, p. 307). Therefore, to evaluate the expected cost associated with a given sequence $(s'_j)_{j=1}^J$, we need to compute N_j, I'_j, and B. These can be computed from K_j, as shown below:

$$I'_j = [s'_j - K_j]^+, \qquad B_j = [K_j - s'_j]^+, \qquad (2)$$

$$N_1 = K_1, \quad \text{and} \quad N_j = K_j - B_{j-1}, \quad j > 1. \qquad (3)$$

However, as noted by LZ, the exact evaluation of the distributions of K_js in the tandem supply network is possible only in two special cases. When $s'_j = 0$, $j = 1, \ldots, J-1$, the supply system becomes an instance of the Jackson-type network (Walrand 1988, p. 41). Note that s'_J does not play a role in determining queue occupancy levels in the network. The joint distribution of queue occupancy levels in a Jackson network has a product-form structure, which means that it can be obtained as the product of marginal queue length distributions. In essence, this means that individual queues in a Jackson network behave as if they are operated independently. Similarly, if $s'_j \to \infty$ for each j, then each stage decouples into an independent $M/M/1$ queue. The major difficulty in evaluating the joint equilibrium distribution of queue occupancy levels at different stages in systems that do not fall into one of the two cases identified above stems from the

fact that there is a positive dependence between the arrival of input units from one stage to the next (see Lee and Zipkin 1992 for details).

Lee and Zipkin use the observations noted above to develop a tractable approximation scheme in which they model the supply system at each stage as an exogenous sequential system. The term *exogenous* is used to underscore their assumption that the overall workload of the supply system is not affected by the demands generated by the supply chain whose performance we are trying to evaluate. This is realistic when the supply system is owned by a large supplier that has many customers, one of which is the supply chain being investigated. The assumption of an exogenous supply system preserves the sequential nature of processing, but the supply lead time is independently and exponentially distributed with parameter $\nu_j = \mu_j - \lambda$.

BPS characterize the distribution of interarrival times to the second stage in a two-stage system. After observing that the coefficient of variation of the interarrival times is between 0.8 and 1, they propose an approximation that assumes that the arrival process at each stage is a Poisson process. Note that both BPS and LZ approximations lead to the same model in the end in which a stage j supply system is assumed to behave like an $M/M/1$ queue in isolation. In effect, these approximations replace a network with $s_j' > 0$ by a network with $s_j' = 0$, $\forall\, j < J$ for the purpose of finding joint queue occupancy levels. Both BPS and LZ report that the approximation works well after comparing approximate performance measures to their estimates obtained via simulation.

BPS also develop, as a possible improvement to the $M/M/1$ (i.e., BPS-LZ) approximation, an alternative approximation based on the use of a $GI/M/1$ queueing model to describe congestion at Stage 2 in a two-stage system. This modified procedure, which makes use of the characterized distribution of interarrival times, is still an approximation because it does not take into account the serial correlation among Stage 2 arrivals. Through numerical experiments, they show that their method approximates the probability of no shipment delay (fill rate) more closely to the simulated values as compared with the BPS-LZ approximation.

Next, we describe a computational procedure, based on the matrix-geometric approach, that can find near-exact performance measures of a series supply system with a small number of stages. We begin by treating the queue at each stage as a finite queue, but one of sufficiently large size that the desired performance measures are quite accurate. We adapt this approach here to develop a procedure to compute the joint distribution of $\mathbf{N}$. The key idea is that the queue occupancy in the stages can be modeled as a quasi-birth-death (QBD) process, which allows us to develop a matrix-geometric solution for its joint distribution.

2.1. Joint Distribution of Queue Occupancy Levels

Let the capacity of stage j queue be q_j, $j = 1, \ldots, J$. Then, the vector $\mathbf{N} = (N_1, N_2, \ldots, N_J)$ represents states of a continuous-time Markov chain with state space $\{\mathbf{n} = (n_1, n_2, \ldots, n_J) \mid 0 \le n_j \le q_j, 1 \le j \le J\}$. Consider an ordering $(\mathbf{m}_0, \mathbf{m}_1, \ldots, \mathbf{m}_{q_1})$ of states, where $\mathbf{m}_k$ is the collection of all states when there are exactly k units in Stage 1, i.e.,

$$\begin{aligned}\mathbf{m}_k = \{&(k,0,\ldots,0,0),(k,0,\ldots,0,1),\ldots,(k,0,\ldots,0,q_J),\\ &(k,0,\ldots,1,0),(k,0,\ldots,1,1),\ldots,\\ &(k,0,\ldots,1,q_J),\ldots,(k,q_2,\ldots,q_{J-1},0),\\ &(k,q_2,\ldots,q_{J-1},1),\ldots,(k,q_2,\ldots,q_{J-1},q_J)\},\end{aligned}$$

is a $\prod_{j=2}^{J}(q_j+1)$-dimensional vector. In this notation, we can represent the stationary distribution of $\mathbf{N}$ by

$$\mathbf{p} = (\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_{q_1}), \tag{4}$$

where the elements of $\mathbf{p}_k$ are the stationary probabilities of states in $\mathbf{m}_k$. Because $\mathbf{N}$ can be regarded as a QBD process, the infinitesimal generator of the chain $\mathbf{N}$ is a block-tridiagonal matrix as given below:

$$Q = \begin{pmatrix} A_0 & B_0 & & & & & & & & \\ C_0 & A_1 & B_0 & & & & & & & \\ & \ddots & & & & & & & & \\ & & C_0 & A_1 & B_0 & & & & & \\ & & & C_0 & A_2 & B_1 & & & & \\ & & & & C_1 & A_3 & B_1 & & & \\ & & & & & & \ddots & & & \\ & & & & & & & C_1 & A_3 & B_1 \\ & & & & & & & & C_1 & A_4 \end{pmatrix}, \tag{5}$$

where A_0, A_1, A_2, A_3, A_4, B_0, B_1, C_0, and C_1 are square matrices of order $\prod_{j=2}^{J}(q_j+1)$. These matrices can be constructed by noting down transition rate-balance equations among the states of the Markov chain. Generally speaking, they are quite sparse. As an illustration, for a two-stage system, we have

$$A_0 = \begin{pmatrix} -\lambda & & & & & \\ \mu_2 & -\lambda-\mu_2 & & & & \\ & \mu_2 & \ddots & & & \\ & & & -\lambda-\mu_2 & & \\ & & & \mu_2 & -\mu_2 & \end{pmatrix}, \qquad B_0 = \begin{pmatrix} 0 & \lambda & & & \\ & 0 & \lambda & & \\ & & \ddots & \ddots & \\ & & & 0 & \lambda \\ & & & & 0 \end{pmatrix}, \tag{6}$$

$A_1 = A_0 - \mu_1 I$, $A_2 = A_1 - \lambda I_{(q_2+1)}$, $A_3 = A_2 + \mu_1 I_{(q_2+1)}$, $A_4 = A_3 + \lambda I$, $B_1 = \lambda I$, $C_0 = \mu_1 I$, and $C_1 = (\mu_1/\lambda)B_0$. $I_{(r)}$ denotes an $r \times r$ matrix of zeroes, except that its last diagonal entry is a 1 and I denotes that its the identity matrix. Thus, we have the steady-state balance equations in matrix form:

$$\begin{aligned} \mathbf{p}_0 A_0 + \mathbf{p}_1 C_0 &= \mathbf{0}, \\ \mathbf{p}_{k-1}B_0 + \mathbf{p}_k A_1 + \mathbf{p}_{k+1}C_0 &= \mathbf{0}, \quad 1 \le k \le s_1' - 1, \\ \mathbf{p}_{s_1'-1}B_0 + \mathbf{p}_{s_1'}A_2 + \mathbf{p}_{s_1'+1}C_0 &= \mathbf{0}, \\ \mathbf{p}_{k-1}B_1 + \mathbf{p}_k A_3 + \mathbf{p}_{k+1}C_1 &= \mathbf{0}, \quad s_1' + 1 \le k \le q_1 - 1, \\ \mathbf{p}_{q_1-1}B_1 + \mathbf{p}_{q_1}A_4 &= \mathbf{0}. \end{aligned} \tag{7}$$

Equations in (7) can be written in the recursive form

$$\mathbf{p}_k = \mathbf{p}_{k-1}R_k, \quad k = 1, 2, 3, \ldots, q_1, \tag{8}$$

where R_k can be found as

$$\begin{aligned} R_{q_1} &= -B_1 A_4^{-1}, \\ R_k &= -B_1[A_3 + R_{k+1}C_1]^{-1}, \quad q_1 - 1 \ge k \ge s_1' + 1, \\ R_{s_1'} &= -B_0[A_2 + R_{s_1'+1}C_1]^{-1}, \\ R_k &= -B_0[A_1 + R_{k+1}C_0]^{-1}, \quad s_1' - 1 \ge k \ge 1, \\ R_0 &= I. \end{aligned} \tag{9}$$

The probabilities $\mathbf{p}_0$ are determined from

$$\mathbf{p}_0(A_0 + R_1 C_0) = \mathbf{0}, \tag{10}$$

subject to the normalization equation

$$\sum_{k=0}^{q_1}\mathbf{p}_k\mathbf{e} = \mathbf{p}_0\sum_{k=0}^{q_1}\left(\prod_{i=0}^{k}R_k\right)\mathbf{e} = 1, \tag{11}$$

where $\mathbf{e}$ is the column vector of ones. This matrix-geometric procedure is very efficient for obtaining the near-exact performance measures through judicious choice of q_js. It also serves as a benchmark for testing approximations and bounds. However, the computational efforts grow rapidly with J and the procedure is time consuming for $J > 3$, making it necessary to investigate approximate procedures.

3. The Approximation

In this section, we briefly explain the BPS approximation, which makes use of the characterization of the input process to Stage 2 in a two-stage system, and we develop an improved approximation scheme. BPS prove that the Laplace-Stieltjes transform of the Stage 2 interarrival time distribution when $s_1' > 0$ can be written as

$$A_2^*(z) = \frac{\lambda}{z+\lambda} - \frac{\rho_1^{s_1'}(\mu_1 - \lambda)z^2}{(z+\lambda)(z+\mu_1)(z+\lambda+\mu_1)}, \tag{12}$$

which upon inversion gives the following complementary cumulative distribution function of A_2 (see Appendix B for details):

$$\bar{F}(y) = e^{-\lambda y} - \rho_1^{s_1'+1}e^{-\lambda y} + \rho_1^{s_1'-1}e^{-\mu_1 y} - \rho_1^{s_1'-1}(1-\rho_1^2)e^{-(\lambda+\mu_1)y}, \quad \forall\, y \ge 0. \tag{13}$$

When $s_1' = 0$, BPS note that $\bar{F}(y) = e^{-\lambda y}$, and therefore Stage 2 is an $M/M/1$ queue. Note that successive interarrival times are serially correlated due to their dependence on Stage 1 queue occupancy levels. However, ignoring the serial correlation, the number of units at Stage 2 at an arbitrary instant of time can be calculated from the standard analysis of $GI/M/1$ queues (see, for example, Kleinrock 1976, p. 252). We denote this approximate distribution as N_2^{BPS} because this is precisely the distribution of N_2 under the BPS approximation.

$$P[N_2^{\mathrm{BPS}} = j] = \begin{cases} 1-\rho_2, & \text{if } j = 0, \\ \rho_2(1-\sigma_2')\sigma_2'^{\,j-1}, & \text{otherwise,} \end{cases} \tag{14}$$

where σ_2' is the solution of the equation $A_2^*(\mu_2-\mu_2 x) = x$. Continuing to use N_2^{BPS} to approximate N_2, we can derive other performance measures of interest. For example, using (3), we obtain the following approximate expression for EK_2 (denoted as EK_2^{BPS}):

$$EK_2^{\text{BPS}} = \frac{\rho_1^{s_1'+1}}{1-\rho_1} + \frac{\rho_2}{1-\sigma_2'}. \tag{15}$$

In contrast, under the BPS-LZ approximation, which assumes that Stage 2 behaves like an $M/M/1$ queue, we will obtain

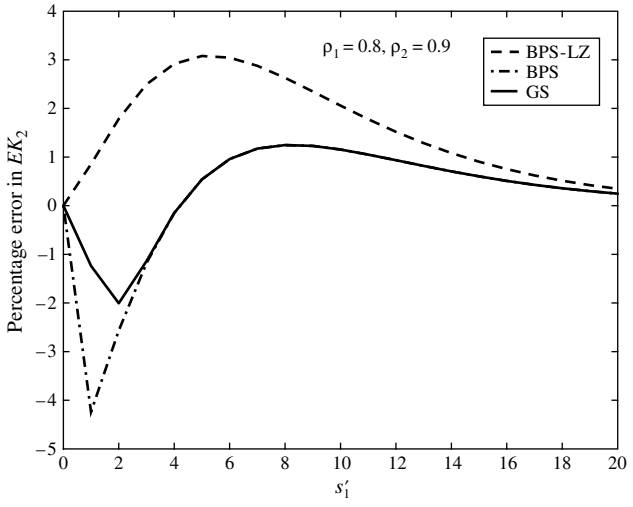
$$EK_2^{\text{BPS-LZ}} = \frac{\rho_1^{s_1'+1}}{1-\rho_1} + \frac{\rho_2}{1-\rho_2}. \tag{16}$$

Numerical comparisons show that EK_2^{BPS} underestimates EK_2 for smaller values of s_1'. When $s_1' = 0$, both the approximations are exact. It can also be seen from (12) that when $s_1' \to \infty$, $A_2^*(z) \to \lambda/(z+\lambda)$, and hence $\sigma_2' \to \rho_2$. Therefore, in that case as well, both approximations converge to the exact value.

However, estimation errors can be large under both approximations when s_1' is positive but small (see, for example, Figure 2). This is in fact a commonly encountered value of s_1' after performing parameter optimization (see §4 for details). To overcome the problem of underestimation when s_1' is small, we propose a modification in which σ_2' is replaced by the following weighted average of σ_2' and ρ_2:

$$\sigma_2 = \left(1 - e^{-s_1'^2/2}\right)\sigma_2' + e^{-s_1'^2/2}\rho_2. \tag{17}$$

Figure 2 **Percentage Errors in the Estimates of EK_2 Under the Three Approximations for the System with $\rho_1 = 0.84$, $\rho_2 = 0.9$ as a Function of s_1'**



The rationale for using σ_2 instead of σ_2' is two-fold. First, the distribution of queue occupancy levels coincides with the exact distribution in the two asymptotic cases: $s_1' = 0$ and $s_1' \to \infty$. Second, it places a greater weight on ρ_2 only when s_1' is very small; its effect diminishes very quickly as s_1' increases. Such a pattern has been observed in numerical experiments carried out with the matrix-geometric method described in §2. The weighting function $e^{-s_1'^2/2}$ is chosen because it is simple and is observed to perform better, on average, than a host of other weights we tested in our numerical experiments. Thus, under the proposed approximation (called the GS approximation), the average number of units that need to be processed at Stage 2's supply system is approximated as

$$EK_2^{\text{GS}} = \frac{\rho_1^{s_1'+1}}{1-\rho_1} + \frac{\rho_2}{1-\sigma_2}, \tag{18}$$

and the simplified expressions for the average number of units in each store can be written as

$$EI_1'(s_1') = s_1' - \frac{\rho_1\left(1-\rho_1^{s_1'}\right)}{1-\rho_1}, \tag{19}$$

$$EI_2'(s_1', s_2') \approx s_2' - \frac{\rho_2\left(1-\sigma_2^{s_2'}\right)}{1-\sigma_2} - \rho_1^{s_1'+1}\left\{\frac{1-\rho_1^{s_2'}}{1-\rho_1} - \rho_2 \sum_{r=0}^{s_2'-1} \sigma_2^r \rho_1^{s_2'-1-r}\right\}. \tag{20}$$

Similarly, the average number of backorders outstanding can be derived to be

$$EB(s_1', s_2') \approx EI_2'(s_1', s_2') - s_2' + EK_2(s_1') = EI_2'(s_1', s_2') - s_2' + \frac{\rho_1^{s_1'+1}}{1-\rho_1} + \frac{\rho_2}{1-\sigma_2}. \tag{21}$$

Under the BPS-LZ and BPS approximations, the performance measures obtained in (19)–(21) have similar expressions after replacing σ_2 by ρ_2 and by σ_2', respectively.

Turning next to supply systems with $J > 2$ stages, we note that the characterization of the arrival process to stage j, with $j > 2$, is difficult because it depends on the number of units that are in the supply systems of the first $(j-1)$ stages. In fact, as the stage index increases it becomes increasingly more difficult to characterize its input process. We therefore adopt

the following decomposition strategy. We assume that stage $(j+1)$ interarrival times depend only on stage j and not on stages $i < j$. This is tantamount to assuming that the local base-stock level at stage $(j-1)$ is large and the supply chain is decoupled at stage j. In such a situation, stage j behaves like the first stage of a two-stage system. Thus, under this approximation, the input process to stage $(j+1)$ is given as

$$A^*_{j+1}(z) = \frac{\lambda}{z+\lambda} - \frac{\rho_j^{s'_j}(\mu_j - \lambda)z^2}{(z+\lambda)(z+\mu_j)(z+\lambda+\mu_j)}. \quad (22)$$

We compute σ'_{j+1} as the solution of the equation $A^*_{j+1}(\mu_{j+1} - \mu_{j+1}x) = x$ and take the weighted average of σ'_{j+1} and ρ_{j+1} to obtain σ_{j+1} as follows

$$\sigma_{j+1} = \left(1 - e^{-s'^2_j/2}\right)\sigma'_{j+1} + e^{-s'^2_j/2}\rho_{j+1}.$$

Here too, we use only the base-stock level at stage j for determining the relative weights of σ'_{j+1} and ρ_{j+1} because s'_j has the greatest effect on the input process to the next stage. The distribution of N_{j+1} is computed with this σ_{j+1}, as shown in (14). Once the distribution of N_{j+1} is computed, it is convoluted with B_j to determine K_{j+1}, which in turn determines the distributions of I_{j+1} and B_{j+1}. We apply the same decomposition strategy to obtain the performance measures under the BPS approximation, but in this case we use σ'_j and not σ_j.

How good is the approximation in (18)–(21) for a two-stage system? Does the performance deteriorate when extended to multistage systems as explained above? We answer these questions in the remainder of this section and in §4. We measure the goodness of the approximation in two different ways: by its ability to accurately predict system performance for a fixed set of base-stock levels, and by its ability to identify the optimal base-stock levels. In the following subsections, we are concerned only with the first of the two types of abilities. The latter is studied in §4.

3.1. Stochastic Comparisons

In the first part of our analysis dealing with performance evaluation, we concentrate exclusively on a two-stage supply chain. The following proposition relates σ'_2 to ρ_2.

Proposition 1. *The solution to the equation* $A^*_2(\mu_2 - \mu_2 x) = x$ *is smaller than or equal to* ρ_2, *where* $A^*_2(\cdot)$ *is given by* (12). *That is,* $\sigma'_2 \le \rho_2$.

The proof is given in Appendix A. Using Proposition 1 and from (17), we have $\sigma'_2 \le \sigma_2 \le \rho_2$. This implies that $EK_2^{\text{BPS}} \le EK_2^{\text{GS}} \le EK_2^{\text{BPS-LZ}}$ (from (15), (16), and (18)). Proposition 1 also leads to the following ordering of the expected backorders under the BPS, GS, and BPS-LZ approximations.

Corollary 1. *In the two-stage supply chain,* $EB_2^{BPS} \le EB_2^{GS} \le EB_2^{BPS\text{-}LZ}$.

Proof. Using the facts that $N_2^{\text{BPS-LZ}}$ has the geometric distribution given by $(1-\rho_2)\rho_2^j$ for $j = 0, 1, \ldots,$ and that the distribution of N_2 under BPS and GS approximations is given in Equation (14) with either σ'_2 or σ_2 as the parameter, we can assert as follows:

$$\begin{aligned} P(N_2^{\text{BPS-LZ}} \le n) &= 1 - \rho_2^{n+1} \\ &\le 1 - \rho_2\sigma_2^n = P(N_2^{\text{GS}} \le n) \\ &\le 1 - \rho_2(\sigma'_2)^n = P(N_2^{\text{BPS}} \le n). \end{aligned} \quad (23)$$

These inequalities immediately imply that $N_2^{\text{BPS}} \le_{\text{st}} N_2^{\text{GS}} \le_{\text{st}} N_2^{\text{BPS-LZ}}$, where $\le_{\text{st}}$ is the usual stochastic order (see Müller and Stoyan 2002, §1.2, for details). Moreover, from the definition of the usual stochastic order, we also have

$$E[\phi(N_2^{\text{BPS}})] \le E[\phi(N_2^{\text{GS}})] \le E[\phi(N_2^{\text{BPS-LZ}})], \quad (24)$$

for any increasing function $\phi(\cdot)$ for which the expectations exist. Next, recall that $B_2 = (N_2 + B_1 - s'_2)^+$ is an increasing function of N_2 and that B_1 is independent of N_2 for all three approximating systems. Therefore, it immediately follows that

$$E[B_2^{\text{BPS}}] \le E[B_2^{\text{GS}}] \le E[B_2^{\text{BPS-LZ}}]. \quad \square \quad (25)$$

Proposition 2. *The distribution of interarrival times to Stage* 2, *distributed according to* (12), *is an NBUE* (*new better than used in expectation*) *distribution*.

Proposition 2, which is proved in Appendix B, implies that A_2 is stochastically less variable than an exponential random variable with the same mean (Ross 1996, p. 437). This result is needed to prove that $EK_2 \le EK_2^{\text{BPS-LZ}}$. BPS have noted this inequality by observing that the coefficient of variation for the interarrival time to Stage 2 is smaller than one. We prove it more formally by making use of some known stochastic comparisons reported in the literature. The result is stated in the following proposition and proved in Appendix C.

Proposition 3. *In a two-stage system, the BPS-LZ approximation provides an upper bound for the true expected value of the number of orders to be processed for demands to date; that is,*

$$EK_2 \le EK_2^{BPS\text{-}LZ}. \tag{26}$$

Thus, the BPS-LZ procedure yields an upper bound on the congestion levels in the supply chain. In contrast, as we shall see in the next section, the BPS and GS methods are not bounds.

Unfortunately, we have not been able to show $EB_2 \le EB_2^{\text{BPS-LZ}}$, although this inequality holds in all numerical experiments that we conducted. Proving a stochastic ordering of backorders is made particularly difficult by two facts: (a) in the original system, the interarrival times are correlated through their dependence on the state of the first stage, and (b) the joint distribution of queue lengths in the two stages can be found only through a numerical procedure. Additional discussion of existing literature on the comparison of $G/GI/1$ queues can be found in Müller and Stoyan (2002, p. 227).

3.2. Numerical Examples

In nearly all numerical experiments reported in this article, we focus on EK_j for some j as the key measure of congestion in the serial supply system. This is deliberate because EI_j' and EB, and therefore total expected cost, are strongly correlated with this statistic. Moreover, for two-stage systems, LZ have reported results of numerical experiments in terms of EK_2 and we use their data to compare the relative accuracy of BPS-LZ, BPS, and GS approximations. In each case, the values of EK_2 are obtained from the three approximation schemes and by the matrix-geometric procedure. The LZ data sets $\lambda = 1$ and uses three different values for μ_j (1.25, 1.5, and 2) corresponding respectively to ρ_j values of 0.8, 0.67, and 0.5. For each of these systems, EK_2 and relative errors are evaluated for five values of s_1', namely, 1, 3, 5, 7, and 9.

Table 1 confirms the facts that, whereas the BPS-LZ approximation overestimates EK_2, the BPS and GS approximations are neither lower nor upper bounds. When s_1' is small, the BPS approximation underestimates EK_2 significantly, and for values of s_1' in the midrange it coincides with the GS approximation. In the midrange, the GS approximation easily outperforms the BPS-LZ approximation. As s_1' becomes larger, all three approximations converge to the true value. It can also be observed that for fixed ρ_1 and ρ_2, the errors initially increase and then begin to decline as s_1' increases (see Figure 2). The mean absolute error for the BPS-LZ approximation is 1.9680%. The same statistic for the BPS approximation is 1.4007% and for the GS approximation is 0.7087%. Similarly, the maximum and minimum absolute errors are, respectively, 3.83% and 0.04% for BPS-LZ, 5.36% and 0.003% for BPS, and 2.13% and 0.01% for GS. These experiments support the claim that the GS approach is a superior approximation scheme.

Table 1 Estimates of EK_2 Under the Three Approximations as Compared with Exact Values

μ_1	μ_2	s_1'	Exact	BPS-LZ$_{approx}$	% error	BPS$_{approx}$	% error	GS$_{approx}$	% error
1.25	1.25	1	7.121	7.200	1.11	6.938	−2.57	7.093	−0.40
1.25	1.25	3	5.866	6.048	3.11	5.879	0.23	5.881	0.26
1.25	1.25	5	5.115	5.311	3.83	5.202	1.72	5.202	1.72
1.25	1.25	7	4.670	4.839	3.62	4.769	2.13	4.769	2.13
1.25	1.25	9	4.405	4.537	2.99	4.492	1.98	4.492	1.98
1.50	1.25	1	5.229	5.333	2.00	4.994	−4.48	5.193	−0.69
1.50	1.25	3	4.440	4.593	3.44	4.440	0.00	4.442	0.04
1.50	1.25	5	4.158	4.263	2.53	4.195	0.89	4.195	0.89
1.50	1.25	7	4.058	4.117	1.46	4.087	0.71	4.087	0.71
1.50	1.25	9	4.021	4.052	0.76	4.039	0.42	4.039	0.42
2.00	1.25	1	4.400	4.500	2.28	4.164	−5.36	4.361	−0.89
2.00	1.25	3	4.059	4.125	1.63	4.040	−0.46	4.041	−0.44
2.00	1.25	5	4.009	4.031	0.56	4.010	0.02	4.010	0.02
2.00	1.25	7	4.001	4.008	0.16	4.002	0.03	4.002	0.03
2.00	1.25	9	4.000	4.002	0.04	4.001	0.01	4.001	0.01

We also report similar results in terms of EB_2 in Table 2. Note that to compute EB_2, we also need to specify s_2'. We have chosen some representative values to highlight the differences between the BPS, GS, and BPS-LZ approximations. These comparisons are qualitatively similar to what we found when comparing EK_2 values. In this series of experiments, the mean absolute errors are found to be 4.93%, 4.65%, and 2.79% with the BPS-LZ, the BPS, and the GS approximations, respectively, over a total of 240 cases.

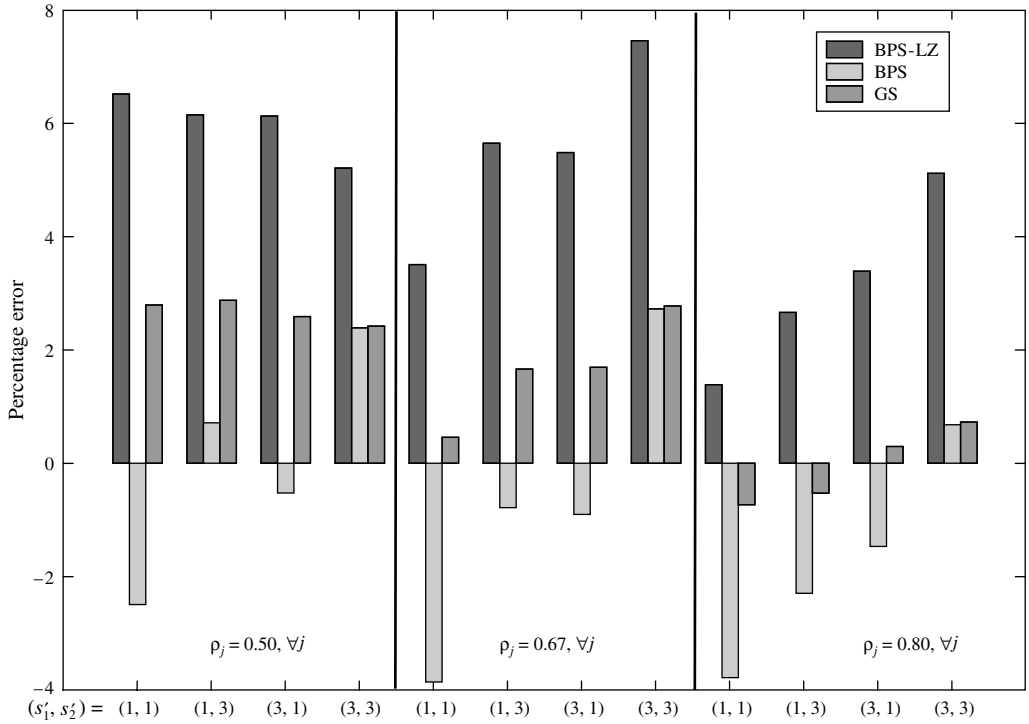
To test the GS approximation in multistage systems, we adapt data from examples reported by Lee and Zipkin (1992) with three stages. We restrict our attention to systems with $\mu_j = \mu$, for all j. Furthermore, we examine three combinations of (μ_1, μ_2, μ_3) with $\mu = 1.25$, 1.5, and 2 and use two values for s_1' and s_2', 1 and 3. The percentage errors in the estimates of EK_3 under the BPS-LZ, BPS, and GS approximations as opposed to the exact values of EK_3 are graphed in Figure 3.

It can be observed from the figure that the observations reported earlier for two-stage systems still hold. When all three approximations overestimate EK_3, the BPS and GS approximations reduce the estimation error by about half in a majority of cases. The mean absolute error for the BPS-LZ approximation is computed to be 4.89%. The same statistic for the BPS and GS approximations are 1.88% and 1.63%, respectively. We have observed this phenomenon in other experiments that have been carried out with different parameter values, but we have not reported them here in the interest of brevity. We have also observed in our numerical experiments that the errors generally increase as the number of stages $J8$ increases. Because

Table 2 Estimates of EB_2 Under the Three Approximations as Compared with Exact Values

μ_1	μ_2	s_1'	s_2'	Exact	BPS-LZ$_{approx}$	% error	BPS$_{approx}$	% error	GS$_{approx}$	% error
1.25	1.25	1	1	6.193	6.272	1.28	6.010	−2.95	6.165	−0.46
1.25	1.25	1	3	4.599	4.669	1.54	4.420	−3.89	4.567	−0.69
1.25	1.25	1	5	3.349	3.408	1.75	3.183	−4.98	3.315	−1.02
1.25	1.25	3	1	4.984	5.166	3.66	4.997	0.27	4.999	0.31
1.25	1.25	3	3	3.557	3.726	4.75	3.568	0.32	3.570	0.37
1.25	1.25	3	5	2.511	2.653	5.65	2.515	0.17	2.517	0.23
1.25	1.25	5	1	4.262	4.458	4.60	4.350	2.06	4.350	2.06
1.25	1.25	5	3	2.937	3.122	6.30	3.022	2.91	3.022	2.91
1.25	1.25	5	5	2.010	2.170	7.94	2.084	3.70	2.084	3.70
1.5	1.25	1	1	4.340	4.444	2.41	4.106	−5.40	4.304	−0.83
1.5	1.25	1	3	2.930	3.018	3.02	2.703	−7.75	2.887	−1.47
1.5	1.25	1	5	1.942	2.009	3.46	1.738	−10.52	1.895	−2.40
1.5	1.25	3	1	3.600	3.753	4.24	3.601	0.00	3.602	0.05
1.5	1.25	3	3	2.341	2.479	5.91	2.340	−0.03	2.342	0.04
1.5	1.25	3	5	1.511	1.621	7.26	1.504	−0.45	1.506	−0.37
1.5	1.25	5	1	3.340	3.446	3.15	3.378	1.11	3.378	1.11
1.5	1.25	5	3	2.141	2.240	4.62	2.178	1.75	2.178	1.75
1.5	1.25	5	5	1.365	1.449	6.12	1.398	2.38	1.398	2.38
2	1.25	1	1	3.550	3.650	2.82	3.314	−6.64	3.511	−1.10
2	1.25	1	3	2.289	2.369	3.45	2.063	−9.90	2.241	−2.11
2	1.25	1	5	1.468	1.524	3.78	1.270	−13.53	1.417	−3.49
2	1.25	3	1	3.246	3.313	2.04	3.227	−0.58	3.228	−0.55
2	1.25	3	3	2.069	2.128	2.87	2.052	−0.83	2.052	−0.79
2	1.25	3	5	1.319	1.364	3.41	1.301	−1.38	1.302	−1.32
2	1.25	5	1	3.206	3.228	0.70	3.207	0.03	3.207	0.03
2	1.25	5	3	2.047	2.068	1.03	2.049	0.09	2.049	0.09
2	1.25	5	5	1.306	1.324	1.36	1.308	0.15	1.308	0.15

Figure 3 **Percentage Errors in EK_3 Under the Three Approximations for the System with Three Sets of ρ_js (0.50, 0.67, and 0.80, $\forall j$) Each with Four Sets of Values for (s_1', s_2') ((1, 1), (1, 3), (3, 1), and (3, 3))**



it is difficult to estimate the exact performance measures for systems with more than three stages, we test the approximations for such systems by comparing estimates from analytical methods and simulation.

We report the results of experiments carried out with a 10-stage system in Figure 4. In all the examples reported, $\mu_j = \mu$, and $s_j' = s'$, $\forall j$. We consider three values for μ and four values for s' with $\lambda = 1$ and compare two statistics, EK_5 and EK_{10}, under the two approximations with the simulated values. The simulation is carried out with the help of Arena software. The simulation is replicated 15 times with each replication length set to 1,000,000 time units. These experiments again confirm that the errors are much smaller with the GS approximation as compared with the BPS-LZ approximation. Also, the GS approximation overcomes, to a great extent, the problem of underestimation of EK_js in BPS approximation when the base-stock levels are small.

4. Optimal Base-Stock Policy

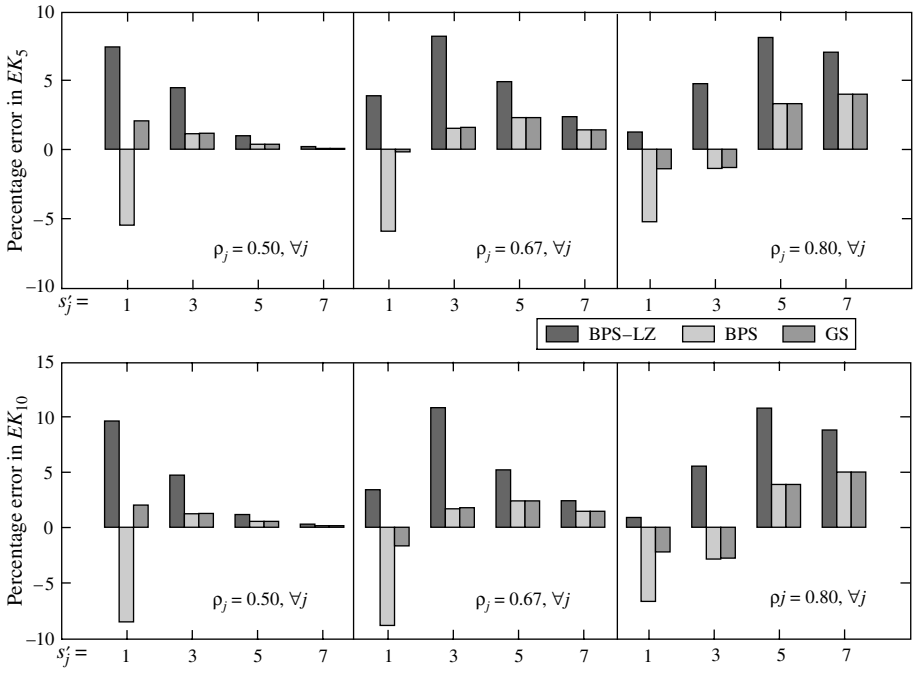
With a good approximation procedure in hand to estimate performance measures of the supply system, we now turn our attention to the optimization of policy parameters. Under the BPS-LZ approximation, EN_j is independent of s_{j-1}' and, as a result, the Clark-Scarf algorithm can be used to determine an optimal echelon base-stock policy, denoted by $\mathbf{s}^*$ (see Zipkin 2000, §8.3.3, p. 308, for details). For a system with J stages, the Clark-Scarf algorithm computes the optimal echelon base-stock policy by minimizing J nested convex functions recursively, starting from stage J. The convexity of the underlying cost functions (denoted as $C_j(\cdot)$ for stage j) holds as long as C_j is independent of s_i, $i < j$. In contrast, under the BPS and GS approximations, EN_j is dependent on s_{j-1}' through σ_j' and σ_j, respectively. Thus, we need to find optimal policy parameters by different means. Consider first a two-stage system for which it is easier to identify patterns of optimal base-stock levels.

4.1. The Two-Stage System

We have two potential stocking points offering different degrees of stockout protection at different costs where the cost function is given by

$$C(\mathbf{s}') = E[h_1'(I_1' + N_2) + h_2' I_2' + bB]. \qquad (27)$$

Figure 4 **Percentage Errors in EK_5 and in EK_{10} Under the Three Approximations for the System with Three Sets of ρ_js, (0.50, 0.67, and 0.80, $\forall j$) Each with Four Sets of Values for s'_js (1, 3, 5, and 7, $\forall j$)**



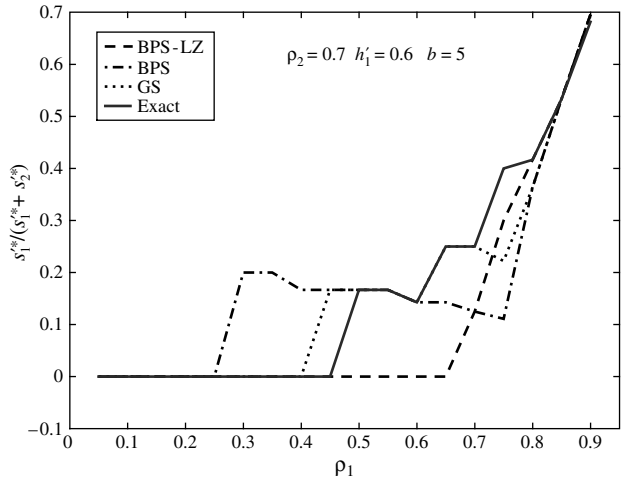
It can be observed that, for each fixed s'_1, the cost function C is convex in s'_2 by virtue of the fact that both EI'_2 and EB are convex in s'_2. Hence, the optimal value of s'_2 for a given s'_1 can be obtained from the usual first-order optimality equations. In the following paragraph, we show that the optimal s'_1 can be obtained by a bounded search, when for each fixed s'_1 the corresponding optimal s'_2 is selected.

Observe that as $s'_1 \to \infty$, $\sigma_2 \to \rho_2$. Therefore we may write $EI'_1(s'_1) = s'_1 - \rho_1/(1-\rho_1) + g_1(s'_1)$, $EI'_2(s'_1, s'_2) = s'_2 - \rho_2(1-\rho_2^{s'_2})/(1-\rho_2) - g_2(s'_1)$, and $EB(s'_1, s'_2) = \rho_2^{s'_2+1}/(1-\rho_2) + g_3(s'_1)$, where $g_i(s'_1)$s are $o(1)$ functions that approach 0 as $s'_1 \to \infty$. Also in the limit as $s'_1 \to \infty$, $EN_2 \to \rho_2/(1-\rho_2)$, which is independent of s'_1. Thus, as s'_1 becomes large, for each fixed s'_2 the backorder cost stabilizes, whereas the inventory costs continue to increase linearly, making the expected total cost monotonically increasing, i.e., it approaches a line with slope h'_1. Clearly, there must exist a finite s'_1 beyond which additional inventory does not reduce the expected total cost. It means that the optimal base-stock level for the first stage lies in some interval, say, $[0, \hat{s}_1]$, where $\hat{s}_1$ is a large constant. In all the numerical experiments conducted for this paper, the value for $\hat{s}_1$ is set at 50. Numerical computations confirm that the cost function C is monotonically increasing when s'_1 exceeds this value.

In what follows we provide some examples to compare the three approximation schemes in terms of their abilities to identify the optimal policy parameters. We fix $\lambda = 1$ (making $\rho_j = 1/\mu_j$) and $h'_2 = 1$ in all our experiments. Three types of experiments are reported: varying first ρ_1 (Figures 5 and 6) and then h'_1 (Table 3), while keeping all other parameters fixed; and varying s'_1 while keeping $s'_1 + s'_2$ fixed at $s'^*_1 + s'^*_2$ (Figure 7), the optimal amount of total inventory. In the first two experiments, each approximation scheme is used to find the optimal base-stock levels and the costs associated with each set of base-stock levels are determined from the matrix-geometric procedure. For the third experiment, only the GS approximation results are reported. The pattern that emerges upon using the other two approximations is identical to what we observe for the GS approximation. We summarize our results first and then explain those patterns that appear to be counterintuitive.

1. For Figures 5 and 6 and Table 3, a total of 28 cases are computed. The GS approximation finds the

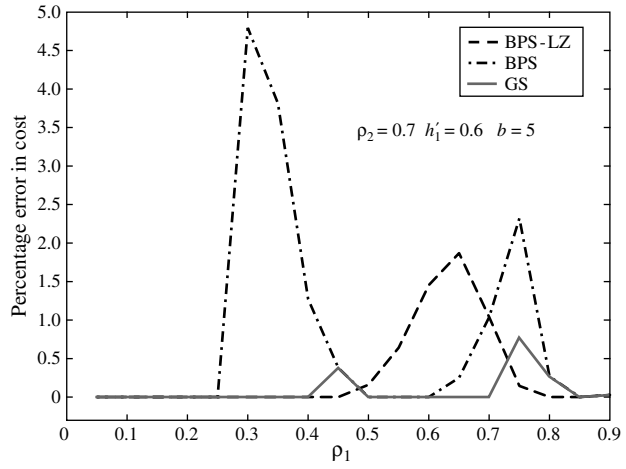
Figure 5 **The Proportion of $s_1'^*$ to $s_1'^* + s_2'^*$ Under the Three Approximations, Along with the Exact Values with $\rho_2 = 0.70$, $h_1' = 0.6$, and $b = 5$**



true optimum in 20 cases, whereas the BPS-LZ and BPS approximations do so in 13 cases each. However, the errors in the optimal costs are higher in the BPS approximation compared with the other two methods. The true optimal base-stock levels and the optimal costs (for all three cases) are calculated using the exact numerical approach outlined in §2. Even when the GS approximation differs from the true optimum values, it overestimates optimal costs by a smaller amount.

2. Base-stock levels obtained from the BPS-LZ approximation agree with a key observation made in Gallego and Zipkin (1999), namely, Stage 1 should hold inventory only either when it is cheaper to hold it there, or when its utilization is much higher relative to Stage 2's utilization. True optimal base-stock levels and the base-stock levels obtained from the BPS and GS approximations do not always agree with this observation.

Figure 6 **Percentage Errors in the Optimal Cost Under the Three Approximations with $\rho_2 = 0.70$, $h_1' = 0.6$, and $b = 5$**



3. Under the BPS and GS approximations, and also under the optimal base-stock policy, we find examples where Stage 1 holds inventory even if its utilization (and hence supply lead time) is not higher than that of Stage 2, and h_1' is only slightly less than h_2'. Moreover, there are examples in which Stage 1 holds inventory when $h_1' = h_2'$ and ρ_1 is only slightly more than ρ_2. In contrast, under the BPS-LZ approximation Stage 1 does not hold inventory in these circumstances.

4. When the amount of inventory at Stage 1 is varied while keeping total system stock level fixed, system cost is found to be quite sensitive to stock allocation. In these experiments, the total stock in the system is fixed at the sum of optimal base-stock levels, $s_1'^* + s_2'^*$, which are in turn computed from the GS approximation. Figure 7 also shows that the system cost rises sharply when Stage 1 holds more inventory at the cost of less-than-optimal amount of stock at Stage 2. Cost penalties are smaller when Stage 2 has more than the optimal amount of stock and when the holding costs at the two stages are nearly equal.

A key difference between the BPS and GS approximations is that in some cases the BPS scheme places more than optimal amount of inventory at Stage 1, which results in a sharp increase in the system cost, whereas the GS approximation minimizes such occurrences resulting in closer to optimal performance. Observations in Item 3 above are a consequence of the congestion effect. By placing inventory at Stage 1, the production system in effect decreases the chance that Stage 2 will starve on account of lack of semifinished parts. Therefore, inventory at Stage 1 increases utilization of Stage 2's capacity, which in turn allows the supply system to achieve similar overall performance with smaller total inventory cost. BPS-LZ approximation treats each stage as an exogenous supply system. Therefore, it does not place any inventory at Stage 1 under the circumstances described in Item 3. The observation in Item 4 stems from the fact that when $h_1' \ll h_2'$, optimal stock levels at Stage 1 are already

Table 3 **Comparison of the Three Approximation Schemes for Different Values of h_1' with $\rho_1 = 0.75$, $\rho_2 = 0.7$, and $b = 7$**

	Exact		BPS-LZ approximation		BPS approximation		GS approximation	
h_1'	$(s_1'^*, s_2'^*)$	Cost	$(s_1'^*, s_2'^*)$	% error in cost	$(s_1'^*, s_2'^*)$	% error in cost	$(s_1'^*, s_2'^*)$	% error in cost
0.10	(9, 6)	6.9844	(10, 6)	0.18	(10, 6)	0.18	(10, 6)	0.18
0.20	(7, 6)	7.7532	(8, 6)	0.17	(8, 6)	0.17	(8, 6)	0.17
0.30	(6, 6)	8.3962	(7, 6)	0.27	(6, 6)	0.00	(6, 6)	0.00
0.40	(6, 6)	8.9737	(5, 7)	0.03	(4, 7)	0.43	(4, 7)	0.43
0.50	(4, 7)	9.4283	(4, 7)	0.00	(4, 7)	0.00	(4, 7)	0.00
0.60	(4, 7)	9.8445	(3, 8)	0.44	(1, 9)	3.10	(3, 8)	0.44
0.70	(3, 8)	10.2358	(2, 9)	0.77	(1, 9)	1.61	(2, 8)	0.47
0.80	(2, 8)	10.5752	(0, 10)	1.84	(1, 9)	0.73	(2, 8)	0.00
0.90	(2, 8)	10.8662	(0, 10)	1.26	(1, 9)	0.34	(2, 8)	0.00
1.00	(1, 9)	11.1549	(0, 10)	0.73	(1, 9)	0.00	(1, 9)	0.00

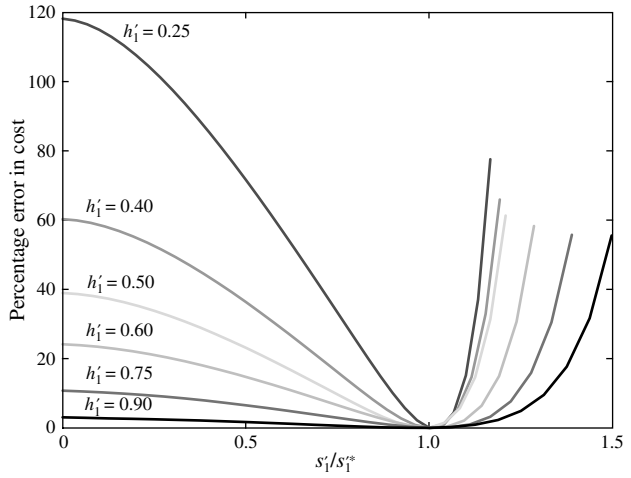
high. Additional allocations in favor of Stage 1 tend to increase shortage costs sharply, while only marginally improving holding costs. Note that if more than the optimal amount of inventory is placed at Stage 1, without a simultaneous decrease in Stage 2 inventory, then the change in system cost may be small under similar parameter values.

4.2. Multistage Systems

With $J > 2$, we can continue to use the Clark-Scarf algorithm for finding optimal stock levels under the BPS-LZ approximation. However, we need to use a bounded search method similar to what was described in the previous subsection when working with the BPS and GS approximations. We solve and report results from a number of numerical experiments involving three stages. We fix $\lambda = 1$, $b = 9$, and assume linear holding costs, i.e., $h_j' = j/J$, in our experiments. The results are presented in Table 4. Because it is computationally demanding to find the true optimal values of s_js, these quantities are not calculated. Instead, approximate best-target levels are found for each of the three approximations and corresponding expected cost is calculated using the matrix-geometric method. The lower the overall cost, the closer the approximation to the least-costly solution. (The smaller values are shown in bold font.)

It is clear that the GS approximation results in a lower cost in a majority of the cases. The performance of the BPS-LZ approximation is superior when the supply systems of the first two stages are severely capacity constrained. This phenomenon can be explained as follows. When a supply system has high-capacity utilization, it is almost always congested and its interdeparture times mirror service times which are exponentially distributed. This gives rise to a Poisson departure process, making the BPS-LZ approximation more accurate. The effect of interstage inventories is relatively small in such cases so long as these inventories are not very large. The BPS approximation is outperformed by the other two approximations. This is due to the fact that the estimates of the performance metrics deviate significantly from the exact values when the base-stock levels are smaller.

Figure 7 **Percentage Error in Cost as a Function of $s_1'/s_1'^*$, While Keeping Total System Stock Equal to the Optimal Total Stock**



Note. Parameters are $\rho_1 = 0.9$, $\rho_2 = 0.6$, $b = 9$ and each curve corresponds to a different value of h_1' in the range 0.25 to 1.0.

Table 4 **Comparison of the Three Approximation Schemes for a Three-Stage System with $\lambda = 1$, $b = 9$, and $h'_j = j/J$**

	BPS-LZ approximation		BPS approximation		GS approximation	
(ρ_1, ρ_2, ρ_3)	(s'^*_1, s'^*_2, s'^*_3)	Cost	(s'^*_1, s'^*_2, s'^*_3)	Cost	(s'^*_1, s'^*_2, s'^*_3)	Cost
(0.50, 0.50, 0.50)	(1, 1, 5)	6.0337	(1, 1, 4)	6.0821	(1, 2, 4)	**5.9395**
(0.50, 0.50, 0.67)	(1, 0, 8)	8.5711	(1, 1, 6)	8.5200	(1, 1, 7)	**8.4643**
(0.50, 0.67, 0.67)	(0, 3, 8)	10.0025	(1, 1, 8)	10.0599	(1, 2, 8)	**9.9467**
(0.67, 0.67, 0.67)	(2, 3, 8)	**10.5949**	(2, 1, 9)	10.7679	(2, 2, 8)	10.6291
(0.70, 0.75, 0.80)	(2, 2, 15)	17.1284	(1, 1, 15)	17.5492	(2, 2, 14)	**17.1052**
(0.50, 0.60, 0.70)	(0, 1, 9)	10.1387	(1, 1, 8)	**9.9941**	(1, 1, 8)	**9.9941**
(0.60, 0.70, 0.80)	(0, 1, 15)	15.9848	(1, 1, 13)	15.8747	(1, 2, 13)	**15.7202**
(0.80, 0.70, 0.60)	(8, 5, 6)	**12.0854**	(8, 5, 6)	**12.0854**	(8, 5, 6)	**12.0854**
(0.50, 0.55, 0.60)	(1, 1, 6)	**7.4844**	(1, 1, 6)	**7.4844**	(1, 1, 6)	**7.4844**
(0.75, 0.75, 0.75)	(3, 4, 12)	15.2204	(2, 1, 14)	15.7019	(3, 3, 12)	**15.2039**
(0.80, 0.80, 0.80)	(5, 5, 15)	**19.7215**	(4, 1, 18)	20.3251	(5, 3, 16)	19.8477
(0.70, 0.70, 0.70)	(3, 3, 9)	**12.0667**	(1, 1, 11)	12.7130	(2, 2, 10)	12.2491
(0.55, 0.70, 0.70)	(0, 3, 9)	11.4382	(1, 1, 9)	11.6897	(1, 2, 9)	**11.3948**
(0.70, 0.55, 0.70)	(4, 0, 9)	10.7141	(4, 1, 8)	**10.6095**	(4, 1, 8)	**10.6095**
(0.70, 0.70, 0.55)	(2, 6, 5)	9.9038	(1, 6, 5)	10.0777	(3, 5, 5)	**9.8077**

5. Parametric Insights

In this section, we use the GS approximation to gain insights for systems with an arbitrary number of stages. Specifically, we will attempt to develop guidelines for which stage(s) to target for holding cost rate reduction when there exists an opportunity to do so by managerial action, and which stage(s) to target for lowering supply lead time by investing in additional capacity. These issues have been studied by Shang and Song (2003a) for systems with constant lead times. In order to allow comparisons to be drawn between their results and ours, we too consider balanced systems, i.e., systems in which each stage has the same utilization. Shang and Song note that as the echelon inventory holding cost h_j increases, the echelon base-stock inventory levels s_i increases for stages $i > j$ but decreases for $i \leq j$. The total expected cost always increases. They also show that reducing the echelon holding cost rate at the farthest upstream stage minimizes optimal cost. Similarly, the best stage to target for lead time reduction is the farthest downstream stage.

We consider a four-stage system to study these issues. First, we consider a system with parameters $\lambda = 1$, $\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0.6$, and $b = 9$. The echelon holding costs for the base case are $h_1 = h_2 = h_3 = h_4 = 2.5$. We reduce the echelon holding costs at Stages 1, 2, 3, and 4 one at a time from 2.5 to 0.25. In each case, the optimal stock levels and cost are calculated using the GS approximation. The results are shown in Table 5. As expected, the maximum reduction in optimal cost occurs when holding cost rate is lowered at Stage 1, because that amounts to reducing the local holding costs at all downstream stages. More importantly, the echelon base-stock level s_j increases when the reduction is at stage j. However, the other optimal echelon base-stock levels do not remain stable, contrary to what is found by Shang and Song (2003a) in constant supply lead time systems.

Often the holding cost rate reductions are the result of local efforts and do not have supply-chain-wide effect. In that case, it makes sense to study the effect on system costs when local holding cost rates are reduced. This is studied in the context of a four-stage system with $\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0.75$, and $b = 15$ in Table 6. In the base case, the local holding costs are $h'_1 = 1$, $h'_2 = 2$, $h'_3 = 3$, $h'_4 = 4$. We then consider a 20%

Table 5 **Effect of Reducing Echelon Holding Cost Rate in a Four-Stage System**

Utilization	(s_1, s_2, s_3, s_4)	Cost	% reduction
Base	(7, 6, 5, 4)	62.2657	—
$h_1 = 0.25$	(11, 6, 5, 3)	42.5408	31.7
$h_2 = 0.25$	(8, 7, 5, 3)	47.8625	23.1
$h_3 = 0.25$	(8, 7, 6, 4)	51.9807	16.5
$h_4 = 0.25$	(8, 7, 6, 5)	56.1438	9.8

Note. The parameters are $\lambda = 1$, $\rho_j = 0.6$, $\forall j$, and $b = 9$ with the base case echelon holding costs being $h_j = 2.5$, $\forall j$.

Table 6 Effect of Reducing Local Holding Cost Rate in a Four-Stage System

Utilization	(s_1', s_2', s_3', s_4')	Cost	% reduction
Base	(3, 3, 2, 11)	60.0887	—
$h_1' = 0.8$	(3, 3, 2, 11)	59.2603	1.39
$h_2' = 1.6$	(2, 4, 2, 11)	58.5508	2.56
$h_3' = 2.4$	(2, 2, 6, 9)	57.8577	3.71
$h_4' = 3.2$	(2, 2, 2, 13)	54.4923	9.31

Note. The parameters are: $\lambda = 1$, $\rho_j = 0.75$, $\forall j$, and $b = 15$ with the base case local holding costs being $h_1' = 1$, $h_2' = 2$, $h_3' = 3$, $h_4' = 4$.

reduction in local holding cost rate applied to each of the four stages, one at a time. In each case, the optimal stock levels and cost are calculated using the GS approximation. Table 6 shows that the maximum reduction in the total average cost occurs when the local holding cost rate is reduced at Stage 4, i.e., at the stage closest to the customer. Even though the holding cost rate reduction can be spread over several stages at a time, it is observed that because for a majority of the parameter values the farthest downstream stage carries more inventories than the other stages, it is beneficial to apply the entire cost reduction to that stage alone. Exceptions do occur, especially when two or more stages carry large inventories.

The observations in the paragraph above apply only to systems that are balanced to begin with and where local holding costs increase linearly with stage index. Clearly, other types of holding cost structures are possible (for example, see the description of kink and jump structures below). However, in such cases, each system needs to be studied on a case-by-case basis to identify where holding cost reduction should be applied. Fortunately, our approximation method provides an efficient and accurate method of evaluating different alternatives.

Next, we study the effect of capacity utilization at each stage on system performance and develop some guidelines about where to add capacity when an opportunity exists to reduce lead times by adding capacity at one or more stages. If $s_j' = 0$ for all $j < J$, the model described in this article becomes mathematically equivalent to the model of an asynchronous flow line (see, for example, Buzacott and Shanthikumar 1993, §5.4). For such systems, it is known that an ordering of stages that places a more capacity-constrained stage earlier minimizes the total amount of inventory in the supply system. Similarly, if work could be distributed to each stage at will, each stage should have an equal amount of work relative to its capacity (i.e., $\rho_j = \rho$ for all j) in order to minimize supply system inventories. These issues are further complicated in the system we study by the presence of planned inventories and unequal holding cost rates.

We consider five different structural forms of holding costs h_j', namely constant, linear, affine, kink, and jump cost structures when evaluating where to add capacity. These are the same forms that are studied in Gallego and Zipkin (1999). Holding cost rates of $h_j' = 1$, for all j, and $h_j' = j/J$, represent, respectively, the constant and linear cost cases. For the affine cost structure $h_j' = \alpha + (1-\alpha)j/J$ for some $\alpha \in (0, 1)$. This represents the case when material at source has some positive cost, and the system then adds cost at a constant rate. The kink structural form is piecewise linear with two pieces. When the kink occurs at the stage $\lceil J/2 \rceil$ ($\lceil x \rceil$ denotes integer ceiling of x), for some $\alpha \in (-1, 1)$, the kink holding cost rate function can be written as

$$h_j' = \begin{cases} (1-\alpha)\dfrac{j}{J}, & j \le J/2, \\ \dfrac{1-\alpha}{2} + \dfrac{(1+\alpha)(j-J/2)}{J}, & j > J/2. \end{cases} \tag{28}$$

In the jump form, the cost is incurred at a constant rate except that one stage has a large cost. If the jump occurs at stage $\lceil J/2 \rceil$, and $\alpha \in (0, 1)$ is a constant, then h_j' is given as follows:

$$h_j' = \begin{cases} (1-\alpha)\dfrac{j}{J}, & j \le J/2, \\ \dfrac{1+\alpha}{2} + \dfrac{(1-\alpha)(j-J/2)}{J}, & j > J/2. \end{cases} \tag{29}$$

In the experiments reported in Table 7, a four-stage system with parameters $\lambda = 1$, $\alpha = 0.6$, and $b = 11$ is studied with each of the five different forms of holding costs. The base case is $\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0.80$. Assuming that capacity utilization can be reduced by a factor of 0.2, we consider different cases involving where this reduction is applied. We can reduce utilization at any one stage from 0.8 to 0.6, or the reduction can be applied to a subset of the four stages in

Table 7 Percentage Reduction in Cost for Increase in Production Rate in a Four-Stage System for Different Structural Forms of Holding Costs

$(\rho_1, \rho_2, \rho_3, \rho_4)$	Constant	Linear	Affine	Kink	Jump
Base case	—	—	—	—	—
(0.80, 0.80, 0.80, 0.60)	13.56	19.72	15.08	25.20	21.87
(0.80, 0.80, 0.60, 0.80)	13.45	13.91	13.06	15.15	14.50
(0.80, 0.60, 0.80, 0.80)	13.76	10.07	11.99	7.76	7.41
(0.60, 0.80, 0.80, 0.80)	5.83	6.40	6.31	3.84	3.64
(0.80, 0.80, 0.70, 0.70)	20.01	24.60	20.42	**30.16**	**30.29**
(0.80, 0.70, 0.80, 0.70)	20.05	21.20	19.53	22.59	21.02
(0.70, 0.80, 0.80, 0.70)	15.02	18.55	15.62	19.97	18.57
(0.80, 0.70, 0.70, 0.80)	20.02	18.43	18.91	16.72	15.64
(0.70, 0.80, 0.70, 0.80)	15.07	15.68	15.13	14.93	13.89
(0.70, 0.70, 0.80, 0.80)	15.26	13.10	14.49	8.66	8.32
(0.80, 0.73, 0.73, 0.73)	**23.75**	**24.99**	**23.44**	26.44	26.36
(0.73, 0.80, 0.73, 0.73)	20.08	22.86	20.62	25.26	25.25
(0.73, 0.73, 0.80, 0.73)	20.16	20.37	20.09	19.20	18.21
(0.73, 0.73, 0.73, 0.80)	20.12	18.67	19.66	15.64	14.68
(0.75, 0.75, 0.75, 0.75)	23.40	23.62	23.28	23.29	23.07

Note. The parameters are $\lambda = 1$, $\alpha = 0.6$, $b = 11$, and the base case $\rho_j = 0.8, \forall j$.

equal amounts. Clearly, there are many more ways of applying the available extra capacity to the processing stages, e.g., by allocating capacity unequally. Similarly, there exist examples in which the initial system is not balanced, or where the holding cost does not vary according to one of the five models considered here. We will comment on our attempts to study the behavior of supply systems under more general conditions after we describe the results reported in Table 7, which are close in spirit to the experiments reported in Shang and Song.

In Table 7, the optimal cost is computed with the help of the GS approximation for each case and the percentage reduction in cost is noted, with the base case acting as the reference. It can be observed that later stages are better targets for capacity increase, with the maximum benefit being realized when the extra capacity is applied equally to the last three stages under constant, linear, and affine cost structures. Under kink and jump structures, the optimal targets for capacity increase are the stages that follow the kink or the jump, in this case Stages 3 and 4. We have investigated (but not reported here) numerous other examples that start with a balanced initial system. The observation that capacity increase should be applied to the stages following kink or jump holds in all cases. For constant, linear, and affine holding cost rate structures we find that the optimal capacity allocations create two subsystems. Each stage within a subsystem has the same utilization, and stages in the later subsystem have smaller utilizations. However, the number of stages in each subsystem are not necessarily the same under different holding cost structures.

When the supply system is not balanced, it is much more difficult to identify dominant strategies for applying additional capacity. Moreover, without reasonable restrictions on how capacity can be added, there are an enormous number of cases that need to be evaluated. Therefore, we restricted attention to experiments in which (a) initially existing capacity is never reduced at a processing stage, and (b) additional capacity is allocated to any subset of processing stages so as to equalize utilization at these stages. This approach is consistent with the way many managers think about adding capacity. Also, it permits unequal capacity allocations because the initial system is not necessarily balanced. In the interest of brevity, we summarize our findings without presenting the experimental data. We find that under kink/jump cost structures, extra capacity is allocated only to stages following the kink/jump and in a manner in which each such stage has the same capacity utilization. Similarly, under the linear holding cost, it is optimal to allocate additional capacity to the last few stages (the number of affected stages varies) in a manner that equalizes capacity utilizations at those stages. Finally, when the holding cost is either constant or affine, the optimal allocation can be either to make the entire system balanced, or to create two balanced subsystems with the later occurring subsystem having smaller utilization.

Having compared and contrasted the three approximation schemes, we close this section with some general guidelines about which method should be used when. We found the BPS approximation to be the least accurate for policy parameter optimization. If the user wishes to estimate costs for a given set of base-stock levels, then the GS approximation is the clear choice. It is considerably better and involves only slightly more computational work. If the goal is rough-cut optimization, then the BPS-LZ approximation may be appropriate, especially when local holding cost rates are different. However, GS approximation provides

a closer-to-optimal allocation of stock among different stages, and leads to significantly lower cost. Even though more effort is needed to find the optimal stock levels when the GS approximation is used, its economic benefits can be large.

6. Fill-Rate-Constrained Optimization

Consider the problem of minimizing the average inventory costs subject to a fill-rate constraint. For serial systems, a base-stock policy is not optimal even when lead times are constant (see Axsäter 2003 for a description of the optimal policy under these conditions). The structure of the optimal policy is not known when lead times are random and serially correlated. Therefore, we shall continue to focus on base-stock policies. For any fixed policy (s_1', s_2'), the steady-state average holding cost is given as

$$H(s_1', s_2') = h_1'(EI_1' + EN_2) + h_2' EI_2', \tag{30}$$

and the fill-rate-based service level constraint is

$$\beta(s_1', s_2') \geq \eta, \tag{31}$$

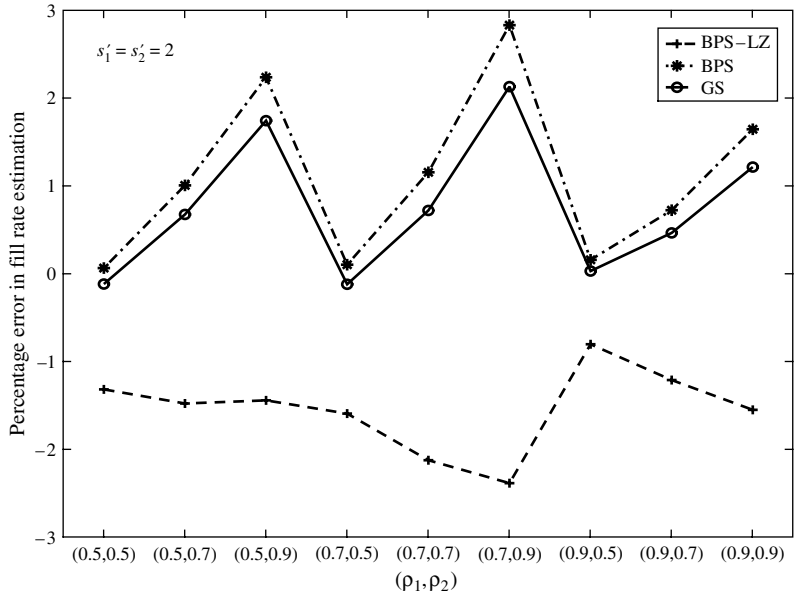
where η is the prespecified fill rate. The simplified expression for $\beta(s_1', s_2')$ under the GS approximation can be written as

$$\begin{aligned}\beta(s_1', s_2') &= P(I_2' > 0)\\ &\approx 1 - \rho_2 \sigma_2^{s_2'-1} - (1-\rho_2)\rho_1^{s_1'+s_2'} - \rho_1^{s_1'+1}\\ &\quad \cdot \rho_2(1-\sigma_2)\sum_{r=0}^{s_2'-2} \sigma_2^r \rho_1^{s_2'-2-r},\end{aligned} \tag{32}$$

where σ_2 is given by (17). Replacing σ_2 respectively by ρ_2 and σ_2', one can obtain the corresponding expressions for fill rate under the BPS-LZ and BPS approximations.

For reasons explained in §3, the BPS-LZ approximation underestimates the true fill rate for a fixed base-stock policy just as it overestimated the congestion levels, whereas the BPS and GS approximations are not bounds. To show the effectiveness of the GS approximation in predicting the fill rate, we consider a two-stage system with $s_1' = s_2' = 2$ and with different values for ρ_1 and ρ_2. The same data have been considered by BPS. The percentage errors in the fill-rate estimation are given in Figure 8. Negative errors are due to underestimation of the true fill rate. It can be observed that the GS approximation provides better estimates of $\beta(s_1', s_2')$ than the other two schemes. The mean absolute error for the BPS-LZ approximation is 1.55%. The same statistic for the BPS and GS approximations is 1.10% and 0.80%, respectively. Similarly, the maximum and minimum absolute errors are, respectively, 2.39% and 0.81% for BPS-LZ, 2.83% and 0.06% for BPS, and 2.13% and 0.03% for GS.

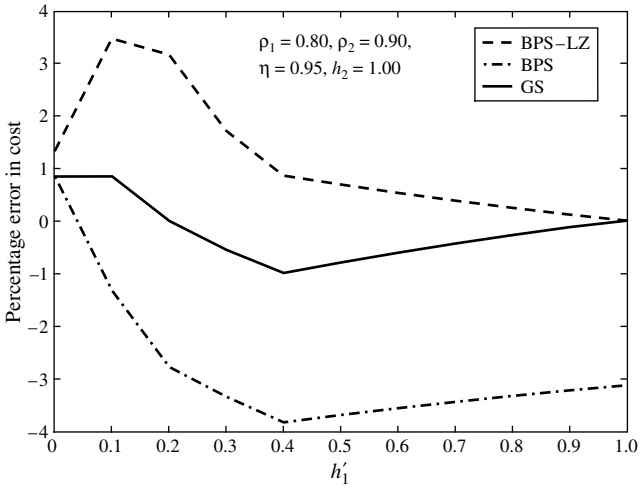
Figure 8 **Percentage Errors in the Estimates of Fill Rate (β) in a Two-Stage System with $s_1' = s_2' = 2$**



Next we turn our attention to the problem of finding the optimal base-stock levels at the two stages for a given set of parameter values and compare the performance of the three heuristics. For this experiment, the parameter values are taken to be $\lambda = 1$, $\rho_1 = 0.8$, $\rho_2 = 0.9$, $\eta = 0.95$, $h_2' = 1$, and we try a range of values for h_1', from 0 to 1. Results are given in Figure 9 and in Table 8. The fill rates and costs for the three heuristics are computed through the matrix-geometric procedure after plugging in the optimal base-stock levels obtained from the approximations. The negative percentage error implies that the fill-rate constraint is not satisfied by the base-stock levels identified as optimal by the approximation. This is primarily a problem with the BPS approximation.

Figure 9 and Table 8 show that the GS approximation estimates are closer to optimal base-stock levels in many cases, whereas there are significant deviations from the true optimum values under the BPS-LZ and BPS approximations. In a number of cases in our experiments, the fill-rate constraint is not satisfied by the optimal stocking levels determined by the BPS approximation. Similarly, the GS approximation does not guarantee that the fill-rate constraint will be satisfied in all cases. From our experiments, it is observed that the optimal base-stock levels estimated under the BPS-LZ approximation are always adequate to satisfy the fill-rate constraint. However, the cost of operating the system with the stocking levels predicted by the BPS-LZ method is higher. We therefore recommend GS approximation for estimating the fill rates, and for fill-rate-based policy optimization.

Figure 9 **Percentage Errors in the Cost Estimates Under the Three Approximation Schemes in a Two-Stage System with $\rho_1 = 0.8$, $\rho_2 = 0.9$, $\eta = 0.95$, and $h_2' = 1$**



7. Concluding Remarks

In this article, we have presented an improved performance evaluation technique for capacitated serial supply systems with planned inventories. Through numerical examples, we have demonstrated that this approach is more accurate when compared with methods proposed in the literature. Moreover, our approach has led to qualitatively different insights. In particular, we have found that upstream stages should hold some inventory even when different stages have nearly identical utilizations and local holding-cost rates. Whereas many managers do hold inventory upstream, this result seems to contradict earlier work that recommends keeping all inventory at the last stage under similar conditions. We use our methodology to study the effect of stock allocation among different stages, when the total amount of stock in the supply system is fixed at the optimal level. We identify conditions under which deviations from optimal allocation lead to significantly poor performance and conditions under which performance is relatively insensitive. We demonstrate how our methodology can be used to determine where to add capacity when each stage of the supply system is a potential candidate. Finally, we show that our scheme continues to work well for a service-constraint-based problem formulation.

Supply managers struggle to determine the best overall stock quantity and its allocation to different stages within multistage supply systems. We have developed an improved procedure for addressing these issues in serial systems. However, there are many related problems that can be pursued in the future. For example, there is a need to develop similar methods for making stocking decisions in capacitated distribution, assembly, and more general supply-system configurations. Another important class of related problems arises when different stages of the supply system are owned by different entities. In these instances, it is important to identify mechanisms for coordinating stocking as well as capacity decisions, while achieving solutions that lead to voluntary compliance from different decision makers.

Table 8 **Comparison of the Three Approximation Schemes in a Two-Stage System for Different Values of h_1' with $\rho_1 = 0.8$, $\rho_2 = 0.9$, $\eta = 0.95$, and $h_2' = 1$**

	Exact		BPS-LZ approximation		BPS approximation		GS approximation	
h_1'	$(s_1'^*, s_2'^*)$	β	$(s_1'^*, s_2'^*)$	β	$(s_1'^*, s_2'^*)$	β	$(s_1'^*, s_2'^*)$	β
0.00	(8, 29)	0.9506	(12, 29)	0.9530	(10, 29)	0.9522	(10, 29)	0.9522
0.10	(8, 29)	0.9506	(7, 30)	0.9544	(2, 31)	0.9448	(5, 30)	0.9506
0.20	(5, 30)	0.9506	(7, 30)	0.9544	(2, 31)	0.9448	(5, 30)	0.9506
0.30	(5, 30)	0.9506	(0, 34)	0.9503	(2, 31)	0.9448	(3, 31)	0.9493
0.40	(2, 32)	0.9503	(0, 34)	0.9503	(2, 31)	0.9448	(3, 31)	0.9493
0.50	(2, 32)	0.9503	(0, 34)	0.9503	(2, 31)	0.9448	(3, 31)	0.9493
0.60	(2, 32)	0.9503	(0, 34)	0.9503	(2, 31)	0.9448	(3, 31)	0.9493
0.70	(2, 32)	0.9503	(0, 34)	0.9503	(2, 31)	0.9448	(3, 31)	0.9493
0.80	(2, 32)	0.9503	(0, 34)	0.9503	(2, 31)	0.9448	(3, 31)	0.9493
0.90	(2, 32)	0.9503	(0, 34)	0.9503	(2, 31)	0.9448	(3, 31)	0.9493
1.00	(1, 33)	0.9505	(0, 34)	0.9503	(1, 32)	0.9450	(2, 32)	0.9503

Acknowledgments

This research has been supported in part by the National Science Foundation (DMII-9988721), and in part by Dofasco Inc. of Hamilton, Ontario, Canada, through research grants to Diwakar Gupta.

Appendix A

Proof of Proposition 1. Let us first recast the equation $A_2^*(\mu_2 - \mu_2 x) = x$ as

$$f(x) - g(x) = 0, \tag{A.1}$$

where, from (12),

$$f(x) = \frac{\lambda}{\lambda + \mu_2 - \mu_2 x} - x, \tag{A.2}$$

and

$$g(x) = \frac{\rho_1^{s_1'}(\mu_1 - \lambda)(\mu_2 - \mu_2 x)^2}{(\lambda + \mu_2 - \mu_2 x)(\mu_1 + \mu_2 - \mu_2 x)(\lambda + \mu_1 + \mu_2 - \mu_2 x)}. \tag{A.3}$$

It is known that ρ_2 is a root of the equation $f(x) = 0$. And, $f(x)$ is monotically decreasing in x for $x \in [0, \rho_2]$ because

$$\frac{df(x)}{dx} = \frac{\rho_2}{(1 + \rho_2 - x)^2} - 1 \le 0, \quad \text{for } x \in [0, \rho_2]. \tag{A.4}$$

It can also be seen that $g(x) \ge 0$ for $x \in [0, 1]$. We are still required to show that $f(0) - g(0) > 0$ in order that $\sigma_2' \le \rho_2$. Recall that σ_2' is the solution of (A.1). We now have

$$f(0) - g(0) = \frac{1}{1 + \rho_2}\left\{1 - \rho_1^{s_1' - 1}(1 - \rho_1)\frac{\mu_2^2}{(\mu_1 + \mu_2)(\lambda + \mu_1 + \mu_2)}\right\} > 0, \quad \text{for } s_1' \ge 1. \qquad \square \tag{A.5}$$

Appendix B

Proof of Proposition 2. By definition (Ross 1996, p. 437), the nonnegative random variable X with a finite mean is called NBUE (new better than used in expectation) if $E[X - a \mid X > a] \le E[X]$ for all $a \ge 0$. If X has a complementary cumulative distribution function $\bar{F}_X$ and $E[X] = \beta$, then this is equivalent to proving

$$\int_a^\infty \frac{\bar{F}_X(y)}{\bar{F}_X(a)}\, dy \le \beta, \quad \forall\, a \ge 0. \tag{B.6}$$

This means that the interarrival time to Stage 2 is NBUE if

$$\int_a^\infty \frac{\bar{F}(y)}{\bar{F}(a)}\, dy \le \frac{1}{\lambda}, \quad \forall\, a \ge 0, \tag{B.7}$$

where $\bar{F}(y)$ is the inverse transform of the function $(1 - A_2^*(z))/z$. Using (12), we can write

$$\begin{aligned}\frac{1 - A_2^*(z)}{z} &= \frac{1}{z} - \frac{1}{z}\left\{(1 - \rho_1^{s_1'+1})\frac{\lambda}{z + \lambda} + \rho_1^{s_1'-1}\frac{\mu_1}{z + \mu_1} - \rho_1^{s_1'-1}(1 - \rho_1^2)\frac{\lambda + \mu_1}{z + \lambda + \mu_1}\right\} \\ &= \frac{1}{z + \lambda} - \rho_1^{s_1'+1}\frac{1}{z + \lambda} + \rho_1^{s_1'-1}\frac{1}{z + \mu_1} - \rho_1^{s_1'-1}(1 - \rho_1^2)\frac{1}{z + \lambda + \mu_1},\end{aligned} \tag{B.8}$$

which, on inversion, gives (13). Let

$$M = \int_a^\infty \bar{F}(y)\, dy - \frac{1}{\lambda}\bar{F}(a).$$

Then, the inequality in (B.7) holds as long as $M \le 0$, which can be seen as follows:

$$M = \rho_1^{s_1'-1} e^{-\mu_1 a}\frac{(1 - \rho_1)}{\lambda}[e^{-\lambda a} - 1] \le 0, \quad \forall\, a \ge 0. \tag{B.9}$$

That is, $A_2(t)$ is an NBUE distribution. $\square$

Appendix C

Proof of Proposition 3. We first state certain known results for a $G/M/1$ queueing system. Let customers arrive

for service in a $G/M/1$ queue at the epochs t_n given by $t_0 = 0$, $t_{n+1} = \alpha_1 + \alpha_2 + \cdots + \alpha_{n+1} = t_n + \alpha_{n+1}$, for any $n = 0, 1, 2, \ldots$. The customer arriving at t_n is called the nth arrival or customer, the time α_n between the $(n-1)$th and nth arrivals is the nth interarrival time. The nth customer requires service for a time γ_n, which is exponentially distributed with mean $1/\mu$. Note that $\{\alpha_n\}$ is a dependent sequence but the service times γ_n are independently and identically distributed (i.i.d.). The sequence $\{\gamma_n\}$ is also independent of the interarrival times. Denote

$$A(t) = P(\alpha_n \le t), \qquad E\alpha_n = 1/\lambda, \tag{C.10}$$

$$B(t) = P(\gamma_n \le t) = 1 - e^{-\mu t}, \quad E\gamma_n = 1/\mu \quad \text{and} \quad \mathrm{Var}(\gamma_n) = 1/\mu^2, \tag{C.11}$$

$$D(t) = P(U_n \le t), \quad \text{where } U_n \equiv \gamma_n - \alpha_n, \tag{C.12}$$

$$D(0) \equiv P(\alpha_n > \gamma_n) = \int_0^\infty (1 - A(t))\mu e^{-\mu t}\, dt, \tag{C.13}$$

$$\rho = \frac{\lambda}{\mu}. \tag{C.14}$$

If the sequence $\{(\alpha_n, \gamma_n)\}$, $n = 0, 1, 2, \ldots$ is metrically transitive (i.e., an ergodic sequence) and if $\rho < 1$, then there exists a stationary sequence $\{w_n\}$ of waiting times satisfying

$$w_{n+1} = [w_n + U_n]^+, \tag{C.15}$$

where w_n is the waiting time of the nth arrival, and converges weakly to w, the stationary waiting time with distribution function W. Also, the mean stationary waiting time is given by

$$Ew = \sum_{n=1}^{\infty} \frac{1}{n} \int_0^\infty x\, dD^{*(n)}(x), \tag{C.16}$$

where $D^{*(n)}$ is the n-fold convolution of D. (See Cohen 1969, Chapter II.5 for details.)

In the sequel, let the Subscripts 1 and 2 stand for $G/M/1$ (original system) and $M/M/1$ (BPS-LZ approximation) queues, respectively. By Proposition 2, the input process to Stage 2 is less variable than an exponential random variable with the same mean. This implies that $A_1 \le_v A_2$, where the subscipt v stands for variability ordering. Because $B_1 =_v B_2$ (an exponential distribution with mean $1/\mu_2$), we can write the following series of inequalities (based on similar comparisons presented in §§5.2, 5.4, and 5.9 of Stoyan 1983):

$$\begin{aligned} D_1 \le_v D_2 &\Rightarrow D_1^{*(n)} \le_v D_2^{*(n)} \\ &\Rightarrow \int_0^\infty x\, D_1^{*(n)}(x) \le \int_0^\infty x\, D_2^{*(n)}(x) \\ &\Rightarrow Ew_1 \le Ew_2, \quad \text{from (C.16)} \\ &\Rightarrow EN_2 \le EN_2^{\text{BPS-LZ}}, \end{aligned} \tag{C.17}$$

where N_2 denotes the number in Stage 2's supply system. This in turn implies that (see (16))

$$EK_2 \le EK_2^{\text{BPS-LZ.}} \qquad \square \tag{C.18}$$

References

Axsäter, S. 2003. Note: Optimal policies for serial inventory systems under fill rate constraints. *Management Sci.* **49** 247–253.

Boyaci, T., G. Gallego. 2001. Serial production/distribution systems under service constraints. *Manufacturing Service Oper. Management* **3** 43–50.

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

Buzacott, J. A., S. M. Price, J. G. 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-Verlag, Berlin, Germany, 445–463.

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.

Cohen, J. W. 1969. *The Single Server Queue*, Vol. 8. North-Holland, Series in Applied Mathematics and Mechanics, American Elsevier Publishing Co., New York.

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

Gallego, G., P. Zipkin. 1999. Stock positioning and performance estimation in serial production-transportation systems. *Manufacturing Service Oper. Management* **1** 77–88.

Kleinrock, L. 1976. *Queueing Systems*, Vol. 1. John Wiley & Sons, New York.

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

Müller, A., D. Stoyan. 2002. *Comparison Methods for Stochastic Models and Risks*. John Wiley and Sons, Ltd., West Sussex, UK.

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

Ross, S. 1996. *Stochastic Processes*, 2nd ed. John Wiley & Sons, New York.

Shang, K. H., J. S. Song. 2003a. Newsvendor bounds and heuristic for optimal policies in serial supply chains. *Management Sci.* **49** 618–638.

Shang, K. H., J. S. Song. 2003b. Analysis of serial supply chains with a service constraint. Working paper, The Fuqua School of Business, Duke University, Durham, NC.

Sobel, M. J. 2004. Fill rates of single-stage and multistage supply systems. *Manufacturing Servie Oper. Management* **6** 41–52.

Stoyan, D. 1983. *Comparison Methods for Queues and Other Stochastic Models*. John Wiley & Sons, Chichester, England.

Walrand, J. 1988. *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, NJ.

Zipkin, P. 2000. *Foundations of Inventory Management*. McGraw-Hill, New York.