

Analysis and Optimization of a Multistage Inventory-Queue System

Liming Liu

Department of Industrial Engineering and Engineering Management, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China, liulim@ust.hk

Xiaoming Liu

Faculty of Business Administration, University of Macau, Macau, China, xmliu@umac.mo

David D. Yao

Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027, yao@columbia.edu

An important issue in the management of supply chains and manufacturing systems is to control inventory costs at different locations throughout the system while satisfying an end-customer service-level requirement. The challenge involved is to solve a nonlinear constrained optimization problem that captures the key dynamics of a complex production-inventory system. In this paper, we first develop a multistage inventory-queue model and a job-queue decomposition approach that evaluates the performance of serial manufacturing and supply systems with inventory control at every stage. We then present an efficient procedure to minimize the overall inventory in the system while meeting the required service level. Our technique is relatively simple and delivers accurate performance estimates. Furthermore, numerical studies generate certain managerial insights into related design and control issues.

Key words: inventory queues; manufacturing systems; supply chains; decomposition; optimal inventory allocation

History: Accepted by Fangruo Chen, former department editor; received November 10, 2000. This paper was with the authors 10 months for 2 revisions.

1. Introduction

In electronics, computer, automobile, and many other industry sectors, a manufacturing and supply system usually takes the form of a complex network of suppliers, fabrication/assembly locations, distribution centers, and customer locations, through which materials, components, products, and information flow (Ettl et al. 2000). Throughout the network, there are different sources of uncertainties associated with supplies (availability, quality, and delivery times), processes (transportation times, machine breakdown, and human performance), and demands (arrival times, batch sizes, and types). These uncertainties and other factors affect the performance of a system, including its service level in terms of fill rate or delivery lead time, which in turn affects the bottom line of an enterprise in today's competitive environment. Among other things, inventories can be used to hedge uncertainties and achieve a specific service level. Because inventory placed at different locations usually incurs different costs and results in different service levels for end customers, the efficient allocation and control of inventory assets presents enor-

mous opportunities and, at the same time, poses a great challenge to many companies.

Motivated by this challenge, in this paper we develop an effective approach to deal with complex supply network design problems involving both queueing delay and stocking control at every node in the network. By modeling the interactions of the queueing delay and stocking control in a network setting, we expand the boundary of the system design methodology.

For a serial supply system, we propose a multistage inventory-queue model. By "inventory queue," we refer to a queueing model that incorporates an inventory control mechanism such as the base-stock control. To evaluate the performance of a multistage system, we decompose it into multiple single-stage inventory queues, each with a modified input (raw material arrival process). Our decomposition approach is computationally simple and provides accurate performance estimates. It also enables us to solve an optimization problem that minimizes the total inventory cost subject to a required service level. Our numerical results reveal a number of insights; some of them are notably different from conclusions

reached in prior studies. For example, we demonstrate that, depending on the cost structure, it may be better to assign less-variable servers to downstream stations instead of upstream stations, as commonly suggested in the literature (for systems in which objectives other than inventory costs are considered). We also demonstrate that by considering the processing delay and inventory holding costs together, there is a definite benefit in managing work-in-process inventory (WIP) actively throughout a supply chain.

The rest of this paper is organized as follows. The related literature is reviewed in §2, followed by model formulation in §3, along with some preliminary results. In §4, we propose a decomposition method that treats the queue length at each stage as an independent sum of a material queue and material backorders (see definitions in §4). Since the material queue and backorders can be readily computed, this decomposition leads to an efficient procedure for network performance evaluation. In §5, we first relax the integer requirement on the base-stock level of the last downstream stage so as to utilize the underlying quasi-unimodal property of the cost function. Based on this property, we construct a recursive optimization procedure to compute the optimal solution of the relaxed multistage problem. The optimal solution to the original problem can be recovered from the solution to the relaxed problem. In §6, based on extensive numerical experiments, we present results that demonstrate the impact of various parameters and provide managerial insights to the design and control of networks of inventory queues. The concluding §7 summarizes the main findings and points out future research opportunities.

2. Literature Review

We are concerned with the performance evaluation and optimization of manufacturing and supply chain systems. In the research literature, queueing-network models are usually used for performance evaluation of multistage discrete manufacturing systems, whereas optimizing inventory control in a network system is commonly associated with multiechelon inventory models. Our problem requires an integration of these two types of models.

Clark and Scarf (1960) consider a multiechelon serial system under periodic review, with constant lead times, unlimited processing capacity, stochastic demands, and a finite decision horizon. This multiechelon inventory optimization problem is decomposed into a set of single-location inventory control problems, and the optimal policy is found to be a modified base-stock policy, i.e., order up to the target echelon base-stock level and ship as much as possible if the entire order cannot be filled. This result has

since stimulated significant research efforts in multiechelon periodic-review systems; refer to the details in the survey articles by Graves (1988) and by Federgruen (1993).

The METRIC model of Sherbrooke (1968) has motivated another important stream of research activities in multiechelon systems under continuous review. While the original work on the METRIC model provides an approximate solution, a number of attempts have since been made to obtain the exact solution, e.g., Axsäter (1990). Svoronos and Zipkin (1991) study continuous-review hierarchical inventory systems with exogenous stochastic replenishment lead times and a one-for-one replenishment policy. By preserving the order of replenishments, the authors are able to approximate the steady-state system performance and to bring out the important role played by the lead-time variance (in contrast to the METRIC model). Refer to Axsäter (1993) for a comprehensive review of multiechelon models under continuous review. Recently, a number of authors have developed models for supply chains based on multiechelon inventory theory. For example, Lee and Billington (1993) use a single-node periodic-review inventory model as a building block to analyze a decentralized supply chain with normally distributed demands and processing lead times. The book edited by Tayur et al. (1999) provides a few more examples of supply chain models.

An extension of the standard periodic-review model is to impose a capacity limit at each stage—the maximum amount of outputs per time unit. Glasserman and Tayur (1994) demonstrate that in a serial system with an echelon base-stock policy, the inventory and backorders are stable if the mean demand per period is less than the capacity at every node. Glasserman (1997) develops bounds and approximations for setting the base-stock levels in the above system. Glasserman and Wang (1998) use a large deviations approach to obtain an asymptotic linear relationship between lead time and inventory as the fill rate approaches 100%.

Buzacott and Shanthikumar (1993) study a multi-cell system, where each cell has a stocking point and the material flow is controlled by a production authorization card (PAC) mechanism. The focus is on deriving bounds and approximations for key performance measures. Other related studies include those on kanban-controlled production lines, e.g., Glasserman and Yao (1994, 1996).

Closely related to our work are two papers by Lee and Zipkin (1992, 1995), where the authors study tandem and distributed production systems with exponential processing times and inventory control at every stage. By assuming that the effective production lead time is equal to the sum of order delay and sojourn time (at the production facility), they

transform the production system into a multiechelon model studied by Svoronos and Zipkin (1991). As such, they are able to use the method from Lee and Zipkin (1992) to obtain system performance measures through approximations involving phase-type distributions. Duri et al. (2000) demonstrate that the approximation method of Lee and Zipkin (1992, 1995) can be extended to systems with general service times with the same phase-type approximation used in Svoronos and Zipkin (1991). The basic structure of the system studied in this paper is similar to that of the systems in Lee and Zipkin (1992, 1995) and Duri et al. (2000). Like those works, we also use a decomposition approach. However, ours is based on a very different idea—decompose the queue at each stage into two components, a backlog queue and a material queue, combined with an effort to characterize the arrival process from the upstream stage (see the sections below for details). Furthermore, we optimize the inventory allocation in the system based on the performance model, whereas prior studies focus entirely on performance evaluation.

Ettl et al. (2000) develop a network of inventory-queue model to analyze complex supply chains. Each stocking location is modeled as an $M^X/G/\infty$ inventory queue operating under a base-stock control policy. By considering the possible delay caused by stock-out and modifying the lead time accordingly, they derive analytical expressions for performance measures and develop a constrained nonlinear optimization model. Like the METRIC model, the work was motivated by industrial applications and has since enjoyed successful implementation. Our study adopts the inventory-service optimization framework of Ettl et al. (2000). Our main focus, however, is to capture the queueing delay at each stage due to limited production capacity, whereas the infinite-server model in Ettl et al. is uncapacitated.

Zipkin (2000) provides a systematic discussion of inventory models with stochastic lead times. Based on the system structure, the models are divided into three groups: exogenous sequential systems, parallel systems, and limited-capacity systems. Exogenous sequential systems (see, for example, Kaplan 1970 and Zipkin 1986) are essentially standard inventory systems with constant lead times replaced by stochastic lead times. In a parallel system, an infinite-server

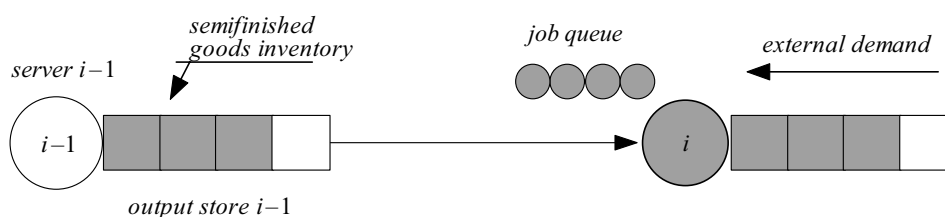
queue is used to model the supply process. With an unlimited capacity, the order lead times are independent and identically distributed random variables. The category of parallel systems (with unlimited capacity) includes a number of interesting works such as Sherbrooke (1968), Berg and Posner (1990), and Ettl et al. (2000). Our model belongs to the third category, models with limited capacity. While we draw heavily on methodologies and results from queueing theory, our model mainly addresses inventory issues. By treating each node of a supply chain as an inventory queue, we emphasize the connection between inventory theory and queueing theory in supply chain applications.

3. The Model and Preliminary Analysis

We consider a manufacturing/supply system with $m + 1$ stages in series (Figure 1), $m \geq 1$, in which each stage consists of a single production/distribution facility. We model this system as a multistage inventory-queue system with $m + 1$ nodes, indexed as $i = 0, 1, \dots, m$. Each node i in the system consists of two parts, a server with service rate μ_i and service time SCV (squared coefficient of variation) C_{si}^2 , and an output store for semifinished products. We assume that the setup/order cost in the system is relatively insignificant, and thus a base-stock policy with a base-stock level $R_i \geq 0$ is used to control the operation of server i and its output store. The output store of node m keeps the finished products to supply external demands and, as such, the system operates in a “make-to-stock” mode.

Whenever an order (demand) arrives at store m , it will be filled immediately if store m has stock on hand. Otherwise, it will be backordered. In either case, a job is added to the job queue—the list of jobs at server m . Concurrently, a job is added to the job queues at server $m - 1$, at server $m - 2, \dots$, all the way to server 0. Thus, the external demand information becomes known to all stations simultaneously, and the production control follows a one-for-one triggering mechanism at all stations along the supply chain. The time to move materials between stages is assumed to be insignificant and hence ignored. Thus, the material movement from an upstream output store to a downstream node also follows the

Figure 1 A Tandem Inventory Queue



one-for-one triggering mechanism in response to the service completion at the downstream node and external demand arrivals. For each node i , we use I_i and B_i to denote the steady-state on-hand inventory level and backorder level, respectively, at its output store, and use N_i to denote the steady-state job-queue length, i.e., the number of outstanding orders in process or waiting to be processed at node i . We note from the above description that an external demand will trigger the increase of N_i and order placements at all nodes. Therefore, the effective demand arrival process at every node is the same as the external demand arrival process. When a server finishes processing a job, the job is stored in the output store (of the node) if there is no backorder; otherwise, it will be used to fill the backorders on a first-come-first-served (FCFS) basis. We further assume that an echelon base-stock policy is used for each node: Server i stops processing when $I_i + \dots + I_m$ reaches $R_i + \dots + R_m$. This policy is a special case of the (Q, r) policy (see, for example, Axsäter and Rosling 1993) with a lot size of one and with the reorder point equal to the base-stock level minus one. From Proposition 1 in Axsäter and Rosling (1993), the echelon base-stock policy for the serial inventory-queue model is equivalent to the installation base-stock policy defined in Axsäter and Rosling. Thus, our inventory-queue model is similar to the standard serial multiechelon inventory models.

External orders (demand) arrive at store m at rate λ , with i.i.d. interarrival times. The traffic intensity at stage i is $\rho_i = \lambda/\mu_i$. Jobs are processed one by one following the FCFS policy with i.i.d. processing times at each node. There is an ample supply of raw materials at the input buffer 0, and all stores are fully stocked initially.

To compute the expected total inventory cost, we classify the WIP in the system into $m+1$ classes. WIP class i ($i=0, 1, \dots, m-1$) refers to those semifinished products (which we shall refer to as product i) that have completed processing at node i , but not yet at node $i+1$ (including the one in process at server $i+1$). Product m is the finished product. Let H_i denote the number of product i , $i=0, 1, \dots, m$. We assume zero replenishment time for raw materials at node 0. Thus, no raw material inventory should be held there.

For a given vector $\mathbf{R} = (R_0, R_1, \dots, R_m)$ of base-stock levels, we are interested in the following steady-state performance measures: the (realized) fill rate f at node m and the expected values of H_i for $0 \leq i \leq m$. These performance measures can be derived from N_i , $i=0, 1, \dots, m$, using the following well-known relations. A demand will be filled immediately if and only if store m has positive on-hand inventory; that is, the number of outstanding orders is less than the base-stock level R_m . Hence, the realized fill rate is

$$f = \mathbf{P}\{N_m < R_m\}. \quad (1)$$

The expected WIPs are given by

$$\mathbf{E}[H_i] = \mathbf{E}[N_{i+1}] + R_i - \mathbf{E}[N_i], \quad i=0, 1, \dots, m-1 \quad (2)$$

and

$$\mathbf{E}[H_m] = \mathbf{E}[R_m - N_m]^+. \quad (3)$$

The following facts and monotone properties are either obvious or readily obtained through simple analysis of the system dynamics.

1. The service completion times and material arrival times at any node are decreasing in the base-stock level at each of its upstream nodes, but are independent of the base-stock level at any other node.

2. The departure times at any node are decreasing in the base-stock level at each of its upstream nodes and the node itself, but are independent of the base-stock level at any other node.

3. The number of outstanding orders at any node is decreasing in the base-stock levels at each of its upstream nodes, but is independent of the base-stock level at any other node.

4. The fill rate at node m is increasing in the base-stock level at every node.

5. $\mathbf{E}[H_i]$, $i=0, 1, \dots, m$, is a nondecreasing function of the base-stock levels at node i and its upstream nodes, but is independent of those at all other nodes.

4. Performance Evaluation

As mentioned above, the exact performance evaluation of the system introduced in the last section is very difficult, even with exponential service times and Poisson demand arrivals. In this section, we develop a decomposition method to approximate the performance. Because all performance measures needed in the optimization model can be derived from job-queue-length distributions, we focus on the approximation of N_i for all $i=0, 1, \dots, m$. Following queueing conventions, we shall use $X/Y/Z/m/R$ to denote a base-stock inventory queue, in which X signifies the external demand process, Y the service process, Z the material supply process, m the number of parallel servers, and R the base-stock level at the output store. Similar to a standard queue, the fundamental process here is the job queue N .

4.1. Job-Queue Decomposition

Node 0 with ample material supplies can be represented as $GI/G/A/1/R_0$, where A signifies an ample supply process. Clearly, the job queue N_0 is identical to that of a standard $GI/G/1$ queue. Node i is a $GI/G/G/1/R_i$ inventory queue, with a material supply process to be determined. The complication here is that a material supply shortage can cause starvation at the downstream stations. One way to account for this is to modify the service time of the downstream

station by adding to it the residual service time at the upstream station. With the revised service time, node i can be analyzed as a standard $GI/G/1$ queue. The advantage of this approach is that the external demand process stays invariant as we analyze each node. However, the starvation probability, which is needed to modify the service times, depends on the operations at both upstream and downstream stations, and hence is difficult to characterize. We have developed and tested some approximations for the starvation probability and found that the accuracy depends heavily on system parameters.

Here, we propose a more direct decomposition/approximation scheme based on the fundamental process N_i . To do so, we need to make a small but important technical modification of how the operation of the system is viewed. Instead of moving a semifinished product to the downstream station for processing when the server becomes available, a semifinished product is moved into the input buffer of the downstream station whenever its job queue is increased by one. When the upstream output store is empty, the request for the semifinished product will be backordered, as shown in Figure 2. Viewed this way, the job queue N_i consists of two parts: a material queue Q_i at the input buffer of node i plus what is on the backorder list U_i (the number of units backordered) at node i .

Hence, we write

$$N_i = Q_i + U_i. \quad (4)$$

Following the common approach in standard queueing networks (also in Lee and Zipkin 1992, 1995), we further assume that Q_i and U_i are probabilistically independent (although they clearly are not). We note that this independence assumption is equivalent to the product-form assumption made in most of the decomposition methods used to analyze queueing networks (refer to, for example, Bitran and Tirupati 1988, Jackson 1963, Kobayashi 1974, Whitt 1984 and the review in Liu 1999).

Liu (1999) has conducted extensive numerical studies with four factors (upstream service-time distribution, upstream base-stock level, downstream service-time distribution, and external demand rate) and a $5 \times 3 \times 5 \times 3$ experiment design to investigate the

impact of the independence assumption. The results demonstrate that the dominating factor is the service-time distribution at the upstream node $i - 1$ (represented by its SCV $C_{s,i-1}^2$), while the impact of the other three factors is very small. In particular, U_i and Q_i tend to be positively correlated when $C_{s,i-1}^2 < 1$, negatively correlated when $C_{s,i-1}^2 > 1$, and the absolute value of the correlation coefficient is smaller than 0.01 when $0.5 \leq C_{s,i-1}^2 \leq 4$. In terms of the impact of $C_{s,i-1}^2$ on the accuracy of the estimation of N_i , the results show that the second and third moments of N_i tend to be underestimated when $C_{s,i-1}^2 < 1$ and overestimated otherwise. Furthermore, the relative error of the second-moment estimate is smaller than 4% for $0.2 \leq C_{s,i-1}^2 \leq 8$. With these observations, we are confident in the independence assumption.

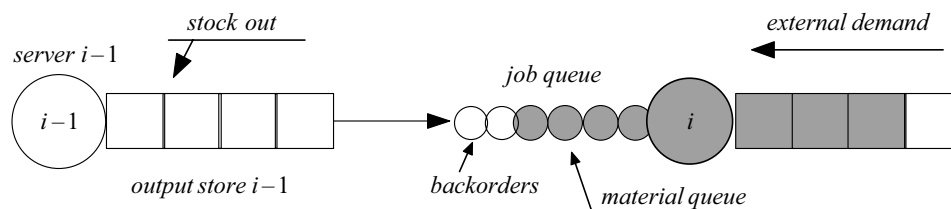
Because node i is the only demand source to the output store of node $i - 1$, we have $U_i = B_{i-1} = [N_{i-1} - R_{i-1}]^+$; hence, the steady-state distribution of U_i can be easily obtained from that of N_{i-1} . Thus, what remains is to derive the probability distribution of Q_i , which is identical to the queue length of a standard $Z/G/1$ queue model, with the input process being the departure process from the output store of node $i - 1$. Also note that $H_i = I_i + Q_{i+1}$.

REMARKS. As pointed out in §2, our decomposition scheme is somewhat similar to the *effective lead-time* decomposition of Lee and Zipkin (1992), with U_i corresponding to the delay and Q_i corresponding to the sojourn time, and both works assume the independence of the two components. The key difference is that Lee and Zipkin (1992) compute the sojourn time independent of the upstream operations, which is equivalent to using the external demand process instead of the departure process from upstream as the input process, to compute the material queue length.

4.2. The Material Queue Q_i

As explained above, we need the departure process from the output store of node $i - 1$, which is the material arrival process at node i , to characterize the material queue Q_i . For multistage systems, the departure process from a node will often be a general (nonrenewal) process, even if the external input process (to the very first node) is Poisson. Thus, it is difficult if not impossible to obtain analytical solutions for the departure process in a multistage

Figure 2 Job Queue as the Sum of the Material Queue and Backorders



tandem inventory-queue system. We have to resort to approximation. In queueing-network literature, a departure process from a node is often approximated by a renewal process, characterized by the first few moments of the interdeparture times (Albin and Kai 1986; Whitt 1982, 1984). This is the approach we adopt here. Specifically, we characterize the departure process from the output store of a general $GI/G/G/1/R$ inventory queue with a base-stock level $R > 0$ by the first two moments of the stationary departure intervals. It is obvious that the mean interdeparture time is equal to the mean interarrival time of external demands. What remains is to approximate the SCV of interdeparture times.

The departure process from an inventory queue is different from the departure process of a standard queueing system. A departure from the output store of node i is either triggered by a service completion at node i when $B_i > 0$ or by an external demand arrival to node $i + 1$ when $I_i > 0$. Our objective here is to provide a simple, yet sufficiently accurate, approximation for this departure process.

Motivated by the widely used approximation of the SCV of the departure intervals from a standard queue (e.g., Buzacott and Shanthikumar 1993),

$$C_d^2 = (1 - \rho^2)C_a^2 + \rho^2 C_s^2, \tag{5}$$

we use the following approximation for the SCV of the departure intervals from output store j , $j = 0, 1, \dots, m - 1$. (See Liu 1999 for numerical validations.)

$$C_{dj}^2 = (1 - \rho_j^{2+R_j/2})C_{aj}^2 + \rho_j^{2+R_j/2}C_{sj}^2, \quad 0 \leq j \leq m - 1, \tag{6}$$

where C_{aj}^2 is the SCV of interarrival times of the material arrival process, and C_{sj}^2 is the SCV of the service times at node j . We note that when $R_j = 0$, (6) is reduced to (5). Intuitively, when R_j is large, departures from the node are more likely responding to the external demands, and hence a larger weight is given to C_{aj}^2 . Thus, adding a function of R_i to the power of ρ_j reflects the impact of the base-stock level on the departure process.

Note that when $C_{aj}^2 = C_{sj}^2 = 1$, the approximation in (6) reduces to $C_{dj}^2 = 1$, which is consistent with the Poisson departure process from a standard $M/M/1$ queue (see Burke 1956). However, the departure process from an $M/M/A/1/R$ inventory queue is not Poisson, as is evident from the following result (refer to Buzacott et al. 1992 and Liu 1999):

$$C_d^2 = 1 - 2\rho^{R+1}(1 - \rho)/(1 + \rho). \tag{7}$$

We choose (6) for its simplicity (so as to be able to handle a large number of nodes) and generality (it applies to nonexponential service times as well). Furthermore, we can show that $1 - 0.2\rho_j^{R_j+1} < C_{dj}^2 \leq 1$

holds in the case of exponential times. Consequently, the error of the approximation in (6) is bounded (see Liu 1999 for details). Numerical evidence regarding the accuracy of the approximation will be discussed in the next subsection.

Having obtained the mean and the SCV of the departure intervals, we can use the following approximation for the material queue-length distribution (Buzacott and Shanthikumar 1993), again for its simplicity and accuracy when applied to our model (see Liu 1999 for details regarding its performance). For $0 \leq i \leq m$,

$$P\{Q_i = j\} = \begin{cases} 1 - \rho_i, & j = 0, \\ \rho_i(1 - \hat{\rho}_i)\hat{\rho}_i^{j-1}, & j \geq 1, \end{cases} \tag{8}$$

where

$$\hat{\rho}_i = \frac{\rho_i(C_{ai}^2 + C_{si}^2)}{\rho_i(C_{ai}^2 + C_{si}^2) + 2(1 - \rho_i)} \tag{9}$$

and $C_{ai}^2 = C_{di-1}^2$.

4.3. Algorithm and Accuracy

We are ready to estimate the overall performance of the inventory-queue network. We start from node 0 and compute U_i , Q_i , and N_i node-by-node until node m . The main steps are summarized below.

Algorithm 1: Performance Evaluation for Multistage Systems.

Step 1. Let $U_0 = 0$ and C_{a0}^2 be the SCV of the interarrival time of external demands. Compute Q_0 with (8) and C_{d0}^2 with (6), $N_0 = Q_0$. Let $i = 0$.

Step 2. Set $i = i + 1$ and $C_{ai}^2 = C_{d,i-1}^2$. Calculate the steady-state distribution of U_i by $U_i = [N_{i-1} - R_{i-1}]^+$ and Q_i by (8). Calculate the steady-state distribution of N_i by $N_i = Q_i + U_i$. If $i = m$, go to Step 4. Else, go to Step 3.

Step 3. Compute the SCV of the departure process C_{di}^2 with (6) and then go to Step 2.

Step 4. Compute the performance measures using relations (1) through (3), and then stop.

Our decomposition method requires only the first two moments of the service times and interarrival times at all nodes. The computational complexity of the procedure is $O((m + 1)K^2)$, where m is the number of stages and K is a constant depending on the required accuracy of the queue-length approximation. We may require, for example, $P(Q_i > K) < 0.01$. Then, $K = \lceil (\ln(0.01) - \ln(\rho_i)) / \ln(\hat{\rho}_i) \rceil$, which increases with $\hat{\rho}_i$. It is easy to derive from (9) that when $C_{ai}^2 + C_{si}^2 < 20$ and $\rho_i < 0.99$, we have $\hat{\rho}_i < 0.95$ and $K \leq 90$. Hence, if we choose $2K$ as a conservative cutoff level when computing Q_i , U_i , and N_i , the computational complexity is $O((m + 1)K^2)$. Note that the computational

complexity is independent of the service-time and interarrival-time distributions.

We now illustrate the overall accuracy of our approximation in terms of $E[H_i]$ ($i = 0, 1, \dots, m$) and the fill rate f (both will be used in the optimization model below). For numerical experiments, we consider a three-stage system (A) with different service-time distributions at different stages and a four-stage system (B) with identical service-time distributions at all stages. We focus on three factors: service-time distributions (*Erlang*, *Exponential* and *HyperExponential* with the SCV less than, equal to, or greater than 1, respectively); the base-stock level at every node; and the external demand arrival rate.

The approximations are compared with estimates from simulation. The results, including the relative errors (err), are presented in Tables 1 and 2 (in which “s” and “a” represent simulation and approximation, respectively). These results show that the approximation works well. We also have the following observations.

OBSERVATION 1. There is no evidence that the error accumulates as we move down the stages.

This is highly desirable and somewhat unexpected (see Table 2). We attribute it to the considerable effort in our decomposition method that is spent on capturing the departure process from upstream. This has isolated the approximation error in approximating the queue length at each stage to that stage only.

OBSERVATION 2. The relative error becomes smaller as the fill rate increases.

Based on the numerical examples, the accuracy is high when the required fill rate is at least 50%. This is likely due to the exponential tail property of the geometric approximation. Indeed, in a separate study (Haque et al. 2002), it is shown through a simple two-stage system that the actual tail behavior of the job-queue length is indeed exponential.

OBSERVATION 3. The approximation is usually better when the base-stock level is higher.

This is consistent with Observation 2 on the fill rate because a higher base-stock level leads to a higher fill rate.

Table 1 Performance Estimations of Multistage System A

$(C_{s0}^2, C_{s1}^2, C_{s2}^2)$	(ρ_0, ρ_1, ρ_2)	Method	$E[H_0]$	$E[H_1]$	$E[H_2]$	f_2
(1,1,1) (2,2,10)	(0.6,0.6,0.6)	s	2.435	2.270	7.866	0.978
		a	2.540	2.350	7.656	0.976
		err%	4.308	3.533	-2.667	-0.262
	(0.9,0.9,0.9)	s	9.086	8.958	1.024	0.212
		a	9.290	9.086	0.920	0.197
		err%	2.244	1.425	-10.171	-6.756
	(0.9,0.8,0.6)	s	4.086	1.624	3.556	0.583
		a	4.290	1.666	3.290	0.556
		err%	4.999	2.595	-7.472	-4.596
s		1.966	9.050	3.235	0.555	
a		2.060	9.549	2.944	0.525	
err%		4.757	5.507	-8.982	-5.294	
(0.6,0.9,0.8)	s	9.924	4.118	3.070	0.534	
	a	10.040	4.236	2.880	0.517	
	err%	1.169	2.866	-6.184	-3.144	
	(0.6,0.6,0.6)	s	2.463	11.926	7.273	0.877
		a	2.527	11.790	7.063	0.889
		err%	2.582	-1.140	-2.883	1.418
(0.9,0.9,0.9)	s	7.597	26.171	2.155	0.324	
	a	7.111	26.097	1.771	0.286	
	err%	-6.393	-0.283	-17.840	-11.605	
(0.25,1,6) (2,10,10)	(0.9,0.8,0.6)	s	3.335	7.895	6.262	0.801
		a	3.450	7.616	6.090	0.815
		err%	3.462	-3.538	-2.745	1.845
	(0.8,0.6,0.9)	s	1.809	31.891	2.886	0.406
		a	1.942	30.303	2.523	0.377
		err%	7.356	-4.980	-12.566	-7.005
(0.6,0.9,0.8)	s	10.094	14.670	3.752	0.525	
	a	9.444	14.837	3.417	0.510	
	err%	-6.444	1.140	-8.932	-2.940	

Table 1 (continued)

$(C_{s0}^2, C_{s1}^2, C_{s2}^2)$	(R_0, R_1, R_2)	(ρ_0, ρ_1, ρ_2)	Method	$E[H_0]$	$E[H_1]$	$E[H_2]$	f_2	
(1,6,0.25)	(0.6,0.6,0.6)		s	12.071	2.967	15.726	0.961	
			a	12.045	2.606	16.054	0.969	
			err%	-0.211	-12.159	2.083	0.826	
	(0.9,0.9,0.9)		s	28.954	18.537	2.161	0.237	
			a	28.335	20.108	2.651	0.263	
			err%	-2.138	8.474	22.712	11.047	
	(10,2,20)	(0.9,0.8,0.6)		s	14.951	2.840	9.786	0.720
				a	14.943	2.629	9.720	0.726
				err%	-0.057	-7.431	-0.676	0.897
(0.8,0.6,0.9)		s	9.920	10.786	9.145	0.783		
		a	9.966	11.001	9.533	0.772		
		err%	0.464	1.995	4.241	-1.316		
(0.6,0.9,0.8)		s	32.729	8.855	5.016	0.448		
		a	32.706	8.606	5.228	0.455		
		err%	-0.069	-2.814	4.227	1.568		
(6,0.25,1)	(0.6,0.6,0.6)		s	8.720	9.608	18.183	0.996	
			a	8.502	9.755	18.227	0.997	
			err%	-2.502	1.527	0.240	0.132	
	(0.9,0.9,0.9)		s	19.344	11.499	4.544	0.395	
			a	17.332	13.496	4.358	0.394	
			err%	-10.397	17.365	-4.103	-0.247	
	(10,10,20)	(0.9,0.8,0.6)		s	10.250	4.133	10.560	0.683
				a	9.046	4.804	10.480	0.700
				err%	-11.748	16.220	-0.763	2.460
(0.8,0.6,0.9)		s	6.686	17.693	9.271	0.721		
		a	6.313	18.656	8.909	0.704		
		err%	-5.575	5.439	-3.908	-2.258		
(0.6,0.9,0.8)		s	14.749	7.770	14.333	0.943		
		a	13.743	8.154	14.833	0.960		
		err%	-6.825	4.948	3.488	1.823		

OBSERVATION 4. The accuracy of performance measures at intermediate nodes seems to be much less sensitive to their base-stock levels than the accuracy of performance measures at the two end nodes.

This is interesting and somewhat unexpected. This may indicate that inventory at the intermediate nodes

has less impact on the fill-rate performance so that changing base-stock levels there results in only small fluctuations in fill rate, and hence small variations in the accuracy of the approximation.

We note that for a two-stage system with exponential service times and a Poisson demand process, our approximation method yields essentially the same numerical results as those in Lee and Zipkin (1992, 1995). We noted in §4.1 the difference between our approximation and that of Lee and Zipkin. Why then do the two methods yield the same numerical results for this example? This has to do with the $M/M/1$ inventory-queue model, which our approximation scheme treats with a Poisson departure process; and this is equivalent to the use of an extraneous sojourn time in Lee and Zipkin (1992).

Table 2 The Relative Errors Over Stages

Service Time	Method	$E[H_0]$	$E[H_1]$	$E[H_2]$	$E[H_3]$	f_3
a	appr.	4.56	4.29	4.18	2.69	0.51
	simu.	4.33	4.08	3.99	3.04	0.55
	err%	5.37	5.19	4.70	-11.56	-6.85
b	appr.	6.18	6.16	6.18	1.51	0.30
	simu.	5.90	5.86	5.85	1.59	0.31
	err%	4.83	5.08	5.58	-4.93	-4.51
c	appr.	2.80	2.31	2.12	6.20	0.92
	simu.	2.62	2.40	2.30	6.25	0.90
	err%	6.72	-3.48	-7.85	-0.79	1.92

Notes. a: Exponential(1.25), SCV = 1.
 b: HyperExponential(0.5,3,0.789474), SCV = 1.78.
 c: Erlang(4,5), SCV = 0.25.
 $\mathbf{R} = (2,2,2,10), \lambda = 1.$

5. Optimization

As demonstrated in the numerical results above, the base-stock level affects end-customer service (fill rate) directly, and its impact differs in location. This, along with the common fact that inventory holding costs at

different locations in a supply chain are usually different, calls for an optimization model so as to set the right base-stock levels in terms of minimizing the overall inventory cost while meeting required end-customer service level.

The optimization problem can be formulated as follows:

$$\begin{aligned}
 OP: \min TC &= \sum_{i=0}^m c_i \mathbf{E}[H_i] \\
 \text{s.t.} \quad &\begin{cases} f(\mathbf{R}) \geq f^r, \\ R_i \in Z^+, \quad i = 0, 1, \dots, m, \end{cases} \quad (10)
 \end{aligned}$$

where f^r denotes the required fill rate, Z^+ the set of nonnegative integers, and $f(\mathbf{R})$ the achieved fill rate given the base-stock vector \mathbf{B} .

OP is a multidimensional discrete nonlinear optimization problem. A direct solution approach is by enumeration, as in Duri et al. (2000), which is obviously impractical when the number of nodes is large. Here, we propose a relaxation-recursive approach, exploiting some special properties of the problem.

Below, we first show that the above objective function exhibits a unimodal property when we relax the integral requirement of the base-stock level at node m . We then construct a set of recursive objective functions that can be optimized sequentially, with the last one in the recursion being the optimization of the relaxed problem. Finally, we explain how to recover a feasible solution to the original problem from the optimal solution to the relaxation problem, and provide an error bound.

For easy reference, we list here the notation to be used in the subsections below.

1. Unit holding costs vector: $\mathbf{c} = (c_0, c_1, \dots, c_m)$.
2. Base-stock vector: $\mathbf{R} = (R_0, R_1, \dots, R_m)$.
3. WIP vector: $\mathbf{H} = (H_0, H_1, \dots, H_m)$.
4. Achieved fill rate: $f(\mathbf{R})$.
5. Upstream base-stock vector: $\tau_i = (R_0, R_1, \dots, R_i)$, $0 \leq i \leq m$.
6. Downstream base-stock vector: $\gamma_i = (R_i, R_{i+1}, \dots, R_m)$, $0 \leq i \leq m$.
7. Invariant upstream base-stock vector: $\tau_i^0 = (R_0^0, R_1^0, \dots, R_i^0)$, $0 \leq i \leq m - 1$. Here the base-stock levels are constants, not decision variables.
8. Downstream partial optimal expected total cost:

$$\begin{aligned}
 h_i(\tau_{i-1}) &= \min_{\gamma_i} \sum_{j=i}^m c_j \mathbf{E}[H_j(\tau_{i-1}, \gamma_i)], \\
 &\quad i = 0, 1, \dots, m, \quad (11)
 \end{aligned}$$

i.e., the minimum total expected holding cost associated with the segment from node i through node m , given the base-stock levels from node 0 through node $i - 1$. (When $i = 0$, the argument of H_j above is

understood to be just γ_0 .) Because the inventory held at stage i does not depend on γ_{i+1} , we have

$$\begin{aligned}
 h_i(\tau_{i-1}) &= \min_{R_i} (\mathbf{E}[H_i(\tau_{i-1}, R_i)] + h_{i+1}(\tau_{i-1}, R_i)), \\
 &\quad i = 0, 1, \dots, m - 1. \quad (12)
 \end{aligned}$$

9. Backward cost functional:

$$\begin{aligned}
 g_j(\tau_{j-1}, R_j) &= c_j \mathbf{E}[H_j(\tau_{j-1}, R_j)] + h_{j+1}(\tau_j), \\
 &\quad j = 1, \dots, m - 1, \quad (13)
 \end{aligned}$$

$$g_0(R_0) = c_0 \mathbf{E}[H_0(R_0)] + h_1(\tau_0), \quad (14)$$

hence

$$h_i(\tau_{i-1}) = \min_{R_i} g_i(\tau_{i-1}, R_i), \quad i = 0, 1, \dots, m - 1. \quad (15)$$

5.1. Relaxation

We consider two balanced systems C and D with exponential service times and a workload of 0.8. System C consists of two stages with a unit cost vector $c = (1, 10)$. System D consists of three stages with a cost vector $c = (1, 2, 3)$. The minimum total inventory holding costs of the two systems as functions of the base-stock levels at node 0, $g_0(R_0)$, are plotted in marked solid curves in Figures 3 and 4, respectively. Clearly, both curves have several local minima. This makes it difficult to design an efficient algorithm to compute the optimal solution.

We believe that the integrality of base-stock levels is the reason why multiple minima exist. Intuitively, because the fill rate and the total inventory holding cost are both increasing in base-stock levels, when the base-stock level at one stage is increased, we should be able to reduce the base-stock level at one of its downstream nodes while maintaining the same fill rate. But this may not always be achieved if the base-stock levels are restricted to integers; for instance, if the required reduction is a fraction.

Therefore, we first relax the integral requirement of R_m . Define the fill rate and the expected on-hand

Figure 3 Cost Function of System C, $c = (1, 10)$

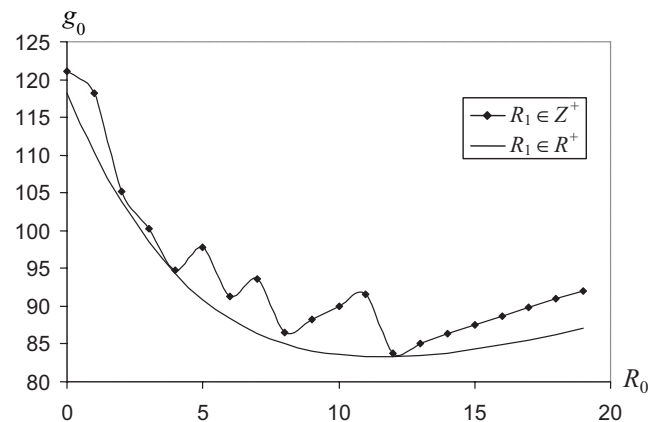
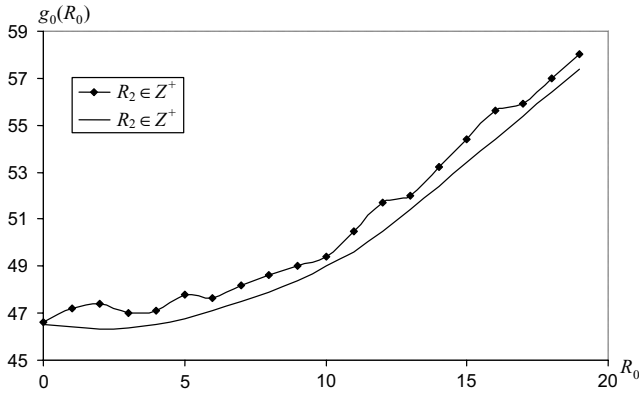


Figure 4 Cost Function of System D, $c = (1, 2, 3)$



inventory corresponding to the real-valued R_m as follows:

$$E[H_i^\pi(\tau_i)] = \begin{cases} E[R_m - N_m(\tau_m)]^+, & i = m, R_m \geq 0, \\ E[H_i(\tau_i)], & i \neq m, \end{cases} \quad (16)$$

$$f^\pi(\mathbf{R}) = f(\tau_{m-1}, \lfloor R_m \rfloor) + (f(\tau_{m-1}, \lceil R_m \rceil) - f(\tau_{m-1}, \lfloor R_m \rfloor))(R_m - \lfloor R_m \rfloor), \quad (17)$$

where $\lfloor x \rfloor$ is the largest integer no greater than x , and $\lceil x \rceil$ is the smallest integer no smaller than x ; the superscript π indicates that the notation (that carries the superscript) corresponds to the real-valued (i.e., relaxed) base-stock levels. Obviously, $E[H_i^\pi(\tau_i)]$ and $f^\pi(\mathbf{R})$ preserve the monotone property of $E[H_i]$ and f , respectively, and have the same value as $E[H_i]$ and f when R_m is an integer.

We now propose the following relaxed optimization problem OP^π :

$$OP^\pi: \min TC^\pi = \sum_{i=0}^m c_i E[H_i^\pi(\tau_i)]$$

$$\text{s.t.} \begin{cases} f^\pi(\mathbf{R}) \geq f^r, \\ R_i \in Z^+, \quad i \neq m, \\ R_m \in R^+. \end{cases} \quad (18)$$

We have tested a large number of different examples and found that for any τ_j , $g_j^\pi(\tau_{j-1}, R_j)$ is always unimodal in R_j for all $j = 0, 1, \dots, m - 1$. Also, it can be seen from Figures 3 and 4 that the global minimum of $g_0^\pi(R_0)$ is very close to the global minimum of $g_0(R_0)$ for both systems.

Hence, we make the following

HYPOTHESIS. For the relaxed optimization problem OP^π , $g_j^\pi(\tau_{j-1}, R_j)$ is unimodal in R_j for all $j \neq m$.

With this hypothesis, we are now in a position to propose a recursive approach to compute the optimal solution of the relaxed problem.

5.2. A Recursive Approach

Let

$$r_m(\tau_{m-1}) = \min\{R_m \in R^+ : f^\pi(\tau_{m-1}, R_m) \geq f^r\}. \quad (19)$$

Because the total cost and the fill rate are both increasing in R_m , for any given base-stock levels at the upstream nodes τ_{m-1} , the optimal cost is achieved when the fill-rate constraint is tight. Thus, we call (19) the last-node optimization. The recursive formulation is then given by

$$h_m^\pi(\tau_{m-1}) = c_m E[r_m(\tau_{m-1}) - N_m(\tau_{m-1})]^+, \quad (20)$$

$$g_k^\pi(\tau_{k-1}, R_k) = c_k E[H_k(\tau_{k-1}, R_k)] + h_{k+1}^\pi(\tau_k), \quad i = 0, 1, \dots, m - 1, \quad (21)$$

$$h_k^\pi(\tau_{k-1}) = \min_{R_k} g_k^\pi(\tau_{k-1}, R_k), \quad k = m - 1, m - 2, \dots, 1, \quad (22)$$

$$h_0^\pi = \min_{R_0} g_0^\pi(R_0). \quad (23)$$

We note that in each recursion k , with the hypothesis made above, $g_k^\pi(\tau_{k-1}, R_k)$ is unimodal in R_j . Thus, the last recursion will give us the optimal solution of the relaxed problem; i.e, $TC^{\pi*} = h_0^\pi$.

An efficient algorithm can be constructed based on the above recursive scheme. Let r be a pointer to a list of integers that represent the base-stock levels; typically, $*(r + k)$ refers to the base-stock level at node k , $k = 0, 1, \dots, m - 1$. Let *Index* be the index for a particular node. A recursive function *OPdown(int Index, int *r)* is the optimal expected total inventory holding cost given that the base-stock levels at the predecessors of node *Index* + 1 are fixed at $(*r, \dots, *(r + \text{Index}))$. The optimization problem OP^π can then be solved by finding the value of *OPdown*(-1, r) using the following C++ program (with the base-stock levels at all nodes initialized at 0).

The Recursive Algorithm.

```

OPdown(double OPdown(int Index, int *r)
{
if (Index == m)
return DealWithLastnode
else
return MinTotalCostGivenTau(OPdown(Index + 1, r)).
}
    
```

DealWithLastnode and *MinTotalCostGivenTau* are two subfunctions. *MinTotalCostGivenTau* returns the minimal total inventory holding cost under the condition that the base-stock levels, at nodes $0, \dots, \text{Index}$, have been fixed at $*r, \dots, *(r + \text{Index})$. Because of the unimodal hypothesis, a golden-ratio search method can

be used in this step. Function *DealWithLastnode* consists of the following steps:

Step 1. Calculate the density of N_0, \dots, N_m , given $\mathbf{R} = (*r, *(r+1), \dots, *(r+m-1), 0)$, using the method described in §4.

Step 2. Obtain the value of r_m^π , such that $f^\pi(*r, *(r+1), \dots, *(r+m-1), r_m^\pi) = f^r$.

Step 3. Calculate the expected total inventory holding cost TC^π and update the best total cost $TC^{\pi*}$ and the best base-stock levels $\mathbf{R}^{\pi*}$ if $TC^\pi < TC^{\pi*}$.

The complexity of this algorithm depends mainly on how many times the function *OPdown* is called. With the unimodal hypothesis, the function in *MinTotalCostGivenTau* is easily evaluated by, say, applying the standard golden-ratio search method. Given that $h(x)$ is unimodal in x and x is an integer, we first identify three points $x_1 < x_2 < x_3$ such that $h(x_2) < \min\{h(x_1), h(x_3)\}$. We then reduce the gap between x_1 and x_3 by the golden ratio ($G = 0.618034$). This is repeated until the gap is less than two. Let the maximum base-stock level at every stage be denoted *MaxR*. The maximum number of iterations needed for each call of the function in *MinTotalCostGivenTau* is $MaxCall = 2 * (\ln 2 - \ln(MaxR)) / \ln(G)$. For example, when $MaxR = 10, 50, 100, 500$, and $1,000$, we have $MaxCall = 8, 14, 18, 24$, and 26 , respectively. (Our numerical experiments show that the actual number of calls needed is generally much lower than *MaxCall*.) For each $Index < m$, *MinTotalCostGivenTau* calls *OpDown(Index + 1, r)* at most *Maxcall* times. Hence, *OpDown(m, r)* is called at most $Maxcall^m$ times.

5.3. Recovering the Optimal Solution

We first present a bound for the difference between the total inventory holding cost of *OP* and its relaxation OP^π , and then introduce a simple method to derive the optimal solution to *OP* from the optimal solution to OP^π . Let $TC^{\pi*}$ and $\mathbf{R}^{\pi*}$ be the optimal (objective) value and the optimal solution to OP^π , respectively.

THEOREM 1. *The optimal values of OP and OP^π satisfy the following inequalities:*

$$\begin{aligned} TC^{\pi*} \leq TC^* \leq TC^{\pi*} + c_m (\lceil R_i^{\pi*} \rceil - R_i^{\pi*}) \\ \leq TC^{\pi*} + c_m. \end{aligned} \quad (24)$$

PROOF. Because OP^π is a relaxation of *OP*, we have $TC^{\pi*} \leq TC^*$. Let \mathbf{R}' be the vector whose components are the ceilings of the corresponding components of $\mathbf{R}^{\pi*}$, i.e., $\tau'_{m-1} = \tau_{m-1}^{\pi*}$ and $R'_m = \lceil R_m^{\pi*} \rceil$. Then,

$$f(\tau'_{m-1}, R'_m) = f^\pi(\tau_{m-1}^{\pi*}, R'_m) \geq f^\pi(\tau_{m-1}^{\pi*}, R_m^{\pi*}) \geq f^r.$$

Hence, \mathbf{R}' is a feasible solution to *OP* and

$$TC^* \leq TC(\mathbf{R}').$$

By definition,

$$\begin{aligned} TC(\mathbf{R}') - TC^\pi(\mathbf{R}^{\pi*}) &= \sum_{i \neq m} c_i \mathbf{E}[H_i(\tau'_{i-1}, R'_i)] + c_m \mathbf{E}[H_m(\tau'_{m-1}, R'_m)] \\ &\quad - \sum_{i \neq m} c_m \mathbf{E}[H_i(\tau_{i-1}^{\pi*}, R_i^{\pi*})] - c_m \mathbf{E}[H_m(\tau_{m-1}^{\pi*}, R_m^{\pi*})] \\ &= c_m \mathbf{E}[H_m(\tau'_{m-1}, R'_m)] - c_m \mathbf{E}[H_m^\pi(\tau_{m-1}^{\pi*}, R_m^{\pi*})] \\ &= c_m (\mathbf{E}[R'_m - N_m(\tau'_{m-1})]^+ - \mathbf{E}[R_m^{\pi*} - N_m(\tau_{m-1}^{\pi*})]^+) \\ &\leq c_m (R'_m - R_m^{\pi*}) \\ &\leq c_m. \quad \square \end{aligned}$$

This theorem shows that we can easily obtain a feasible solution \mathbf{R}' to *OP* from the optimal solution to OP^π , and the difference between the two corresponding objective values is bounded by c_m . When $f(\mathbf{R}') > f^r$, this solution may be further improved by reducing the base-stock levels at upstream nodes while satisfying the fill-rate requirement as follows: If $R_m^{\pi*}$ is an integer, then $\mathbf{R}^* = \mathbf{R}^{\pi*}$ is the optimal solution to *OP*; hence, stop. Otherwise, $\mathbf{R}' = \lceil \mathbf{R}^{\pi*} \rceil$ is feasible to *OP*, but the realized fill rate $f(\mathbf{R}') > f^r$. The base-stock levels at some of the upstream nodes may be further reduced. Starting from the node with the highest unit holding cost (node $m-1$ in this case), reduce its base-stock level one unit at a time until a further reduction will violate the fill-rate constraint. Apply the same procedure to the node with the next-highest unit holding cost. Repeat this until reaching the node with the lowest unit holding cost.

6. Numerical Studies

In this section, we investigate a number of design and control issues using a three-stage tandem system as an example. Our numerical results have confirmed some of the classical and intuitive conclusions, for instance, that the optimal total cost increases in the service-level requirement and the service-time SCV, and that the optimal base-stock level increases in the service-time SCV. We will not discuss these results here; instead, we focus on several new observations.

We note that while the unit holding costs (c_i) here are increasing (in i) as we move downstream, representing a value-added production process, the optimization model *OP* can handle any inventory holding cost structure. Refer to Liu (1999) for a discussion on non-value-added systems.

6.1. The Value of Intermediate Inventory Control

Here, we are interested in understanding the impact of an inventory control policy that sets different inventory targets at different nodes (including all the

intermediate nodes) of a supply system. We consider different combinations of workloads and cost parameters. For each combination, we calculate the optimal total inventory holding cost of the network without intermediate inventory control (called *NOP*), denoted by TC_{wo}^* . This is obtained by letting $R_0 = 0$ and $R_1 = 0$, and finding the minimal value of R_2 such that the service-level requirement is satisfied. We then compare TC_{wo}^* with TC_w^* , the optimal value of *OP*. The results are shown in Table 3. It is obvious that $TC_{wo}^* \geq TC_w^*$ in all cases. Furthermore, the following observations can be made.

OBSERVATION 5. The impact of intermediate inventory control is more significant when the value-added is larger.

When the value-added is high from stage to stage, there is more room for cost savings from optimizing inventory allocation. In this case, a supply chain should be organized such that the high value-added processes are located as close to end customers as possible. Also in this case, we should keep much of the total inventory at upstream nodes so as to reduce the overall holding cost while meeting the service-level requirement.

OBSERVATION 6. The difference between TC_{wo}^* and TC_w^* is more pronounced when node 0 has the highest workload.

This observation may be explained as follows. It is obvious that the node with the highest workload will be likely to have a large number of outstanding orders. Increasing the base-stock levels at its upstream

nodes will not help because this will just increase the congestion at its input buffer. On the other hand, if node 0, which has no upstream nodes, has the highest workload, increasing its base-stock level while reducing those of downstream nodes will result in better inventory allocation and reduced *WIP*. In other words, when node 0 is the bottleneck, the optimal inventory allocation is more effective in reducing the total cost and maintaining a better material flow through the supply chain. This observation also highlights the need to consider workload allocation and inventory allocation simultaneously.

6.2. TC^* and Workload

Here, we investigate how workloads affect the optimal total cost. First, consider system A with $c = (1, 1.5, 2.25)$ and a fill-rate requirement of 0.9. The relationship between the optimal cost and the workload (at every node) is plotted in Figures 5 and 6. From these figures we can make the following observation:

OBSERVATION 7. TC^* is increasing and convex in the workload and is proportional to $1/(1 - \rho)$.

The phenomenon revealed in Figure 6 is quite interesting, but the rationale behind it is not immediately obvious. We may, however, use a single-stage model to investigate this phenomenon. Consider the optimal inventory holding cost of an *M/M/A/1/R* inventory queue with workload ρ and fill-rate requirement f^r . Because the inventory holding cost and the fill rate are increasing in the base-stock level R , the cost

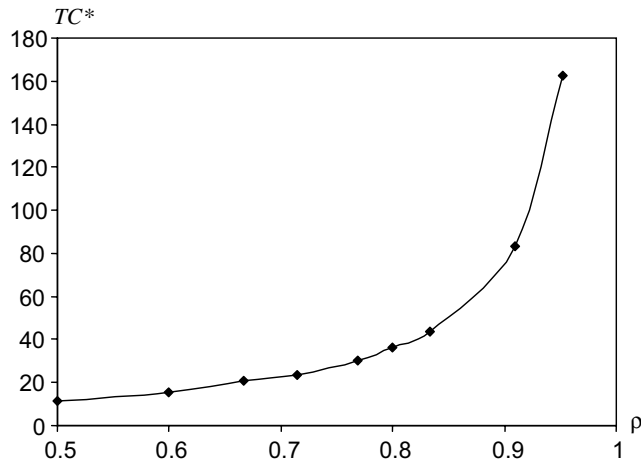
Table 3 TC^* With or Without Intermediate Inventory Control

Case	R_{0wo}^*	R_{1wo}^*	R_{2wo}^*	TC_{wo}^*	R_{0w}^*	R_{1w}^*	R_{2w}^*	TC_w^*	Cost Diff. %
a-1	0	0	10	16.53	0	3	7	15.78	4.74
a-2	0	0	50	76.94	1	19	32	76.47	0.61
a-3	0	0	29	40.92	8	9	12	34.42	18.89
a-4	0	0	29	49.67	0	0	29	49.67	0
a-5	0	0	29	49.67	0	12	18	48.38	2.66
b-1	0	0	10	123.26	4	2	6	96.73	27.42
b-2	0	0	50	539.43	17	19	26	434.29	24.21
b-3	0	0	29	322.76	26	8	7	148.52	117.32
b-4	0	0	29	354.01	11	2	23	338.93	4.45
b-5	0	0	29	339.01	0	19	14	259.46	30.66
c-1	0	0	10	584.46	3	5	5	399.4	46.33
c-2	0	0	50	2,518.42	28	23	24	1,767.72	42.47
c-3	0	0	29	1,559.77	37	9	6	522.59	198.47
c-4	0	0	29	1,632.27	11	2	23	1,516.98	7.6
c-5	0	0	29	1,589.77	1	25	12	995.57	59.68

Notes. $C_{si}^2 = 1, i = 0, 1, 2$ and $f_2^r = 0.9$.

- a: $(c_0, c_1, c_2) = (1.0, 1.5, 1.5^2)$.
- b: $(c_0, c_1, c_2) = (1.0, 4.5, 4.5^2)$.
- c: $(c_0, c_1, c_2) = (1.0, 10, 10^2)$.
- 1: $(\rho_0, \rho_1, \rho_2) = (0.6, 0.6, 0.6)$.
- 2: $(\rho_0, \rho_1, \rho_2) = (0.9, 0.9, 0.9)$.
- 3: $(\rho_0, \rho_1, \rho_2) = (0.9, 0.8, 0.6)$.
- 4: $(\rho_0, \rho_1, \rho_2) = (0.8, 0.6, 0.9)$.
- 5: $(\rho_0, \rho_1, \rho_2) = (0.6, 0.9, 0.8)$.

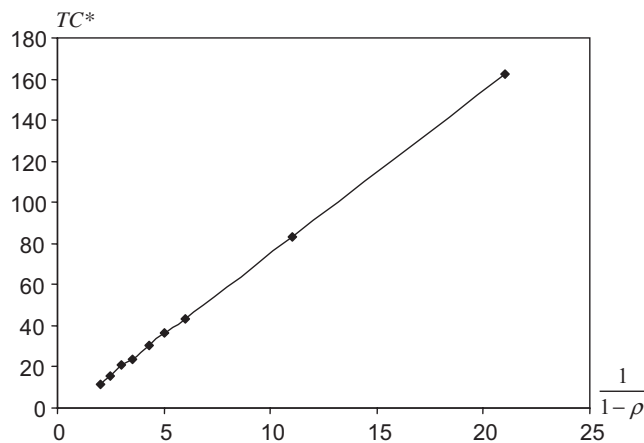
Figure 5 TC^* and Workload I



optimization leads to the minimal base-stock level that meets the fill-rate requirement. Let R^* be the optimal base-stock level. Because $f = (1 - \rho^R)$, we have $f^r \approx (1 - \rho^{R^*})$. Hence, the optimal total inventory is

$$\begin{aligned}
 E[H] &= E[R^* - N]^+ \\
 &= \sum_{i=0}^{R^*-1} (R^* - i)P(N=i) \\
 &= \sum_{i=0}^{R^*-1} (R^* - i)(1-\rho)\rho^i \\
 &= (1-\rho)\rho^{R^*} \sum_{j=1}^{R^*} j\rho^{-j} \\
 &= (1-\rho)\rho^{R^*} \rho^{-1} \\
 &\quad \cdot [1 - \rho^{-R^*}(R^* + 1 - R^*/\rho)] / (1-1/\rho)^2 \\
 &\approx \frac{(1-\rho)(1-f^r)}{\rho} \\
 &\quad \cdot \left(1 - \frac{1}{(1-f^r)}(R^* + 1 - R^*/\rho)\right) / (1-1/\rho)^2
 \end{aligned}$$

Figure 6 TC^* and Workload II



$$\begin{aligned}
 &= R^* - \frac{\rho}{1-\rho} f^r \\
 &\approx \frac{\rho}{(1-\rho)} (-f^r) + \frac{\ln(1-f^r)}{\ln \rho} \\
 &\approx \frac{\rho}{(1-\rho)} (-f^r) - \frac{\ln(1-f^r)}{(1-\rho)}, \\
 &\quad \text{when the workload is close to 1} \\
 &= \frac{(-f^r - \ln(1-f^r))}{(1-\rho)} + f^r. \tag{25}
 \end{aligned}$$

Hence, the optimal total cost is proportional to $1/(1 - \rho)$ for the $M/M/A/1/R$ inventory queue.

Next, we consider the sequencing of workloads. Consider a product that needs to be processed first by station 0, which has a workload of 0.8. The product then needs to be processed at both station 1 and station 2, but the order is not fixed. Suppose all processing times are exponential. Station 1 has workload 0.6 and station 2 has workload 0.9. Let station 0 be assigned to node 0 at the upstream. How should we then sequence station 1 and station 2?

Suppose that the cost parameters are of the form $c = (1, x, x^2)$ with $x > 1$ (we consider only the value-added case, as mentioned earlier). Numerical results shown in Table 4 suggest that:

OBSERVATION 8. It is better to sequence the station with a higher workload first.

At first glance, this seems counterintuitive, or is at least inconsistent with known conclusions. For instance, the “bowl phenomenon” is well known in the workload-allocation literature (see, for example, Hillier and Boling 1979), i.e., for throughput maximization, the machine with the smallest workload should be assigned to the middle of the line. This discrepancy may be explained from two angles. First, the system considered here has an active inventory control at every stage with an objective to minimize the total inventory cost subject to a downstream fill-rate requirement, whereas the “bowl phenomenon” focuses on throughput maximization. As there is a difference in both systems and objectives, it is not unreasonable that the optimal workload arrangements are different. Second, it is perhaps more important to place a higher workload process first at the upstream (which means node 1, as node 0 is taken already) than to place the process with the least workload in the middle. Our results show that it is important to understand the dynamics and priorities among workloads, throughput, inventory costs, and customer service levels. These require further modeling and analysis and more extensive numerical studies.

When the required fill rate is high, we expect the advantage of sequencing a bottleneck or high workload station first becomes more significant. This is

Table 4 TC^* and Workload Sequencing

x	f^r	$\rho = (0.8, 0.6, 0.9)$					$\rho = (0.8, 0.9, 0.6)$					Diff. %
		R_0^*	R_1^*	R_2^*	TC^*	f^*	R_0^*	R_1^*	R_2^*	TC^*	f^*	
1.1	0.6	0	0	15	16.63	0.60	0	0	15	15.88	0.60	4.72
1.1	0.9	0	0	29	30.04	0.90	0	17	13	29.67	0.91	1.27
1.5	0.6	0	0	15	24.73	0.60	0	0	15	20.98	0.60	17.88
1.5	0.9	0	0	29	49.67	0.90	0	21	9	40.68	0.90	22.08
3.0	0.6	5	0	12	66.12	0.60	0	11	5	41.72	0.60	58.48
3.0	0.9	7	0	25	163.40	0.90	2	23	7	97.23	0.90	68.06
5.0	0.6	5	0	12	146.90	0.60	0	13	4	80.86	0.60	81.66
5.0	0.9	13	0	24	410.17	0.90	4	22	7	206.80	0.90	98.34

evident in Table 4, when the fill rate increases from 0.6 to 0.9 for $x = 1.5, 3,$ and 5 . However, when $x = 1$, this is not the case. How do we explain this? As noted above, workloads, fill-rate requirement, and cost structure interact in the system. Obviously, a higher fill-rate requirement will increase the optimal total inventory cost. When the value-added is high, a higher fill-rate requirement will likely cause an even higher increase in the optimal total inventory cost. Thus,

OBSERVATION 9. The advantage of a better workload sequencing is more significant when the value-added is high.

When the value-added is small, the significance of workload sequencing will likely diminish, to the point when another factor becomes dominant. In the case when $x = 1$, the integer requirement of the base-stock level may play a bigger role so as to diminish the impact of better workload sequencing when the fill-rate requirement is high. Nonetheless, a high to low workload sequencing is still better.

6.3. TC^* and Service-Time Variation Sequence

Consider a four-node system. Suppose a product needs to go through node 0 first for an exponential service time with rate 1.25. It then needs to go through each of the three remaining nodes exactly once in any order. Suppose that all three nodes have the same service rate 1.1111, and SCVs 0.25, 1, and 6, respectively. How do we sequence the three nodes?

Suppose the required service level is 0.9 and the cost vector is $\mathbf{c} = (1, x, x^2, x^3)$. We consider two sequences of the nodes in terms of their SCV's, $a: (0.25, 1, 6)$ and $b: (6, 1, 0.25)$. The results are summarized in Table 5, from which the following observations can be drawn.

OBSERVATION 10. When there is no intermediate inventory control in the system, sequence a is better than sequence b .

This is intuitive and consistent with existing results (e.g., Hopp and Spearman 1996), because less variability will be propagated from upstream nodes to

downstream nodes so that both congestion and delay will be lower.

OBSERVATION 11. When intermediate inventory control is allowed, sequence a is better than b for small x values. When x increases beyond a certain point, sequence b becomes better than a .

We know that with a higher SCV there will be more congestion in front of the corresponding process/node. If the value-added is high enough, the cost from the additional WIP at downstream stations may more than offset the benefit from reduced overall system variability when lower SCV processes are placed upstream. This phenomenon cautions us that when intermediate inventory control is present, some of the well-known conclusions and rules may no longer be valid and have to be reevaluated along with the system configuration and the objective function.

7. Concluding Remarks

We have proposed a multistage inventory-queue model for a class of manufacturing and supply systems. Each station in the system is modeled by a single-server queue controlled by a base-stock policy, namely, an inventory queue. A job-queue decomposition scheme is developed to approximate key performance measures, at a level of complexity comparable to that of evaluating single-server queues. Because it is computationally efficient and reliably accurate, the method is suitable for the analysis and optimization of complex supply networks.

In this context, the problem of minimizing inventory costs subject to a service-level constraint is a multidimensional integer optimization problem. We constructed a relaxation of this problem, and proposed a method to obtain a feasible solution to the original problem, which is close to the optimal solution to the relaxed problem. An error bound was also developed. While the solution so obtained is usually very close to the optimal solution, it sometimes results in a service level that is slightly lower than what is required. In this case, a simulation-based method can be deployed to fine-tune the solution so as to enforce the required service level; refer to Liu (1999).

Table 5 TC^* and SCV Sequencing

x	f'_3	Case	Without Planned Intermediate Inventory					With Planned Intermediate Inventory				
			R_{0wo}^*	R_{1wo}^*	R_{2wo}^*	R_{3wo}^*	TC_{wo}^*	R_{0w}^*	R_{1w}^*	R_{2w}^*	R_{3w}^*	TC_w^*
1.0	0.6	a	0	0	0	41	47.225855	0	0	6	34	47.263536
1.0	0.6	b	0	0	0	64	74.744798	1	20	0	33	58.94998
1.1	0.6	a	0	0	0	41	56.493465	0	0	6	34	56.458434
1.1	0.6	b	0	0	0	64	84.681696	1	21	0	32	65.379961
1.2	0.6	a	0	0	0	41	66.958005	0	0	6	34	66.833195
1.2	0.6	b	0	0	0	64	95.910166	1	21	0	32	72.320241
1.5	0.6	a	0	0	0	41	106.188575	0	0	6	34	105.685321
1.5	0.6	b	0	0	0	64	138.356226	0	27	0	28	91.578849
2.0	0.6	a	0	0	0	41	201.628302	0	0	8	33	204.259918
2.0	0.6	b	0	0	0	64	244.369221	0	32	0	25	141.139155
3.0	0.6	a	0	0	0	41	534.707669	0	0	11	30	522.706575
3.0	0.6	b	0	0	0	64	634.161426	0	45	8	14	261.260809
5.0	0.6	a	0	0	0	41	1,999.048327	0	0	15	27	1,857.657657
5.0	0.6	b	0	0	0	64	2,478.741341	0	58	10	11	753.203238
1.0	0.9	a	0	0	0	78	76.528624	0	0	8	69	75.956613
1.0	0.9	b	0	0	0	117	116.288661	1	32	0	58	77.892905
1.1	0.9	a	0	0	0	78	95.495451	0	0	9	68	94.085652
1.1	0.9	b	0	0	0	117	139.976579	1	37	0	53	86.199384
1.2	0.9	a	0	0	0	78	117.59319	0	0	9	68	115.582786
1.2	0.9	b	0	0	0	117	167.697962	2	40	0	50	97.461063
1.5	0.9	a	0	0	0	78	205.085421	0	0	10	67	199.50667
1.5	0.9	b	0	0	0	117	278.566765	2	49	0	44	138.33024
2.0	0.9	a	0	0	0	78	436.050455	0	0	12	65	418.714272
2.0	0.9	b	0	0	0	117	576.720128	1	59	17	25	199.549452
3.0	0.9	a	0	0	0	78	1,325.882435	0	0	13	64	1,252.329449
3.0	0.9	b	0	0	0	117	1,755.845737	1	72	19	21	474.510265
5.0	0.9	a	0	0	0	78	5,661.894465	0	0	20	60	5,203.142894
5.0	0.9	b	0	0	0	117	7,671.724262	3	90	22	18	1,669.911586

Through extensive numerical studies, we have observed a number of interesting properties and gained some useful managerial insights in many aspects of such systems. Some recent studies had to simplify their analyses by, say, not considering queueing processes or cost objectives, leading to findings that raised doubts on the value of active local inventory control. By considering both queueing and inventory aspects of the network, along with a cost objective that emphasizes the inventory-service trade-off, our findings have brought out the value of intermediate inventory control and demonstrated that the specific controls need to be responsive to the cost objective.

The inventory-queue model proposed here can be extended to analyze more complex supply networks. We are currently modifying the model to study distributed (disassembly) systems where one has to consider additional management issues such as stock allocation policies. Another promising direction is the study of the flow time (cycle time) in an inventory-queue model. This will enable us to analyze systems operating under different modes, such as make-to-order and assemble-to-order systems.

Acknowledgments

Two referees and the associate and the area editors have all provided insightful comments and useful suggestions on an

earlier version of this paper. These were very helpful for the authors' improvement of both the content and the exposition of this paper. This study was supported in part by RGC Grant HKUST 6063/97E. The third author was supported in part by NSF Grant DMI-0085124 and RGC Grants CUHK3476/99E and CUHK4173/03E.

References

Axsäter, S. 1990. Simple solution procedures for a class of two-echelon inventory problems. *Oper. Res.* **38** 64–69.

Axsäter, S. 1993. Continuous review policies for multi-level inventory systems with stochastic demand. S. Graves, A. H. G. R. Kan, P. H. Zipkin, eds. *Handbook in Operations Research and Management Science: Logistics of Production and Inventory*. North-Holland, Amsterdam, The Netherlands, 175–198.

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

Albin, S. L., S. R. Kai. 1986. Approximation for the departure process of a queue in a network. *Naval Res. Logist. Quart.* **33** 129–143.

Berg, M., M. Posner. 1990. Customer delay in $M/G/\infty$ repair systems with spares. *Oper. Res.* **38** 344–348.

Bitran, G. R., D. Tirupati. 1988. Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Management Sci.* **35** 851–878.

Burke, P. J. 1956. The output of a queueing system. *Oper. Res.* **4** 699–704.

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. T. Fandel, T. Gullledge, A. Jones, eds. *New Directions for Operations Research in a Manufacturing System*. Springer, 445–463.
- Clark, A., H. Scarf. 1960. Optimal policies for multi-echelon inventory problems. *Management Sci.* **6** 474–490.
- Duri, C., Y. Frein, M. Di Mascolo. 2000. Performance evaluation and design of base-stock systems. *Eur. J. Oper. Res.* **127** 172–188.
- Ettl, M., G. E. Feigin, G. Y. Lin, D. D. Yao. 2000. A supply network model with base-stock control and service requirements. *Oper. Res.* **48** 216–232.
- Federgruen, A. 1993. Centralized planning models for a multi-echelon inventory system under uncertainty. S. Graves, A. H. G. R. Kan, P. H. Zipkin, eds. *Handbook in Operations Research and Management Science: Logistics of Production and Inventory*. North-Holland, Amsterdam, The Netherlands, 133–174.
- Glasserman, P. 1997. Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* **45** 244–257.
- Glasserman, P., S. Tayur. 1994. The stability of a capacitated multi-echelon production-inventory system under a base-stock policy. *Oper. Res.* **42** 913–924.
- Glasserman, P., Y. Wang. 1998. Leadtime-inventory tradeoffs in assemble-to-order systems. *Oper. Res.* **46** 858–871.
- Glasserman, P., D. D. Yao. 1994. *Monotone Structure in Discrete-Event Systems*. Wiley Interscience, New York.
- Glasserman, P., D. D. Yao. 1996. Structured buffer allocation problems. *Discrete Event Dynam. Systems: Theory Appl.* **6** 9–42.
- Graves, S. C. 1988. Safety stocks in manufacturing systems. *J. Manufacturing Oper. Management* **1** 67–101.
- Haque, L., L. Liu, Y. Zhao. 2002. Tail asymptotics of a two-stage inventory-queue model. Working paper, School of Mathematics and Statistics, Carleton University, Ottawa, Canada.
- Hillier, F. S., R. Boling. 1979. On the optimal allocation of work in symmetrically unbalanced production systems with variable operations times. *Management Sci.* **25** 721–728.
- Hopp, W. J., M. L. Spearman. 1996. *Factory Physics*. Irwin, New York.
- Jackson, J. R. 1963. Jobshop-like queuing systems. *Management Sci.* **10** 131–142.
- Kaplan, R. 1970. A dynamic inventory model with stochastic lead times. *Management Sci.* **16** 491–507.
- Kobayashi, H. 1974. Application of the diffusion approximation to queueing networks I: Equilibrium queue distributions. *J. ACM* **21** 316–328.
- Lee, H. L., C. Billington. 1993. Material management in decentralized supply chains. *Oper. Res.* **41** 835–847.
- Lee, Y. J., P. H. Zipkin. 1992. Tandem queues with planned inventories. *Oper. Res.* **40** 936–947.
- Lee, Y. J., P. H. Zipkin. 1995. Processing networks with inventories: Sequential refinement systems. *Oper. Res.* **43** 1025–1036.
- Liu, X. M. 1999. Performance analysis and optimization of supply networks. Ph.D. thesis, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China.
- Sherbrooke, C. C. 1968. METRIC: A multi-echelon technique for recoverable item control. *Oper. Res.* **16** 122–141.
- Svoronos, A., P. H. Zipkin. 1991. Evaluation of one-for-one replenishment policies for multi-echelon inventory systems. *Management Sci.* **37** 68–83.
- Tayur, S., R. Ganeshan, M. Magazine. 1999. *Quantitative Models for Supply Chain Management*. Kluwer Academic Publishers, Norwell, MA.
- Whitt, W. 1982. Approximating a point process by a renewal process, I: Two basic methods. *Oper. Res.* **30** 125–147.
- Whitt, W. 1984. Approximations for departure processes and queues in series. *Naval Res. Logist. Quart.* **31** 499–521.
- Zipkin, P. 1986. Stochastic leadtimes in continuous-time inventory models. *Naval Res. Logist. Quart.* **33** 763–774.
- Zipkin, P. 2000. *Foundations of Inventory Management*. McGraw-Hill, New York.