

A Single-Product Inventory Model for Multiple Demand Classes

Hasan Arslan

Sawyer Business School, Suffolk University, Boston, Massachusetts 02108, harslan@suffolk.edu

Stephen C. Graves

Leaders for Manufacturing Program and A. P. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, sgraves@mit.edu

Thomas A. Roemer

The Rady School of Management, University of California, San Diego, La Jolla, California 92093, troemer@ucsd.edu

We consider a single-product inventory system that serves multiple demand classes, which differ in their shortage costs or service-level requirements. We assume a critical-level control policy, and a backorder clearing mechanism in which we treat a backorder for a lower-priority class equivalent to a reserve-stock shortfall for the higher-priority class. We show the equivalence between this inventory system and a serial inventory system. Based on this equivalence, we develop a model for cost evaluation and optimization under the assumptions of Poisson demand, deterministic replenishment lead time, and a continuous-review (Q, R) policy with rationing. We propose a computationally efficient heuristic and develop a bound on its performance. We provide a numerical experiment to show the effectiveness of the heuristic and the value from a rationing policy. Finally, we describe how to extend the model to permit service times, and how to extend the model to a multi-echelon setting.

Key words: inventory rationing; service differentiation; serial inventory system; priority demand classes

History: Accepted by Paul H. Zipkin, operations and supply chain management; received March 12, 2005. This paper was with the authors 6 months for 2 revisions. Published online in *Articles in Advance* July 25, 2007.

1. Introduction and Literature Review

In many inventory settings, a supply firm wishes to provide different levels of service to different customers. For instance, in a service-parts network, a customer can choose amongst different contracts, each with a different cost and level of service. A “gold contract” might provide a 99% fill rate within 24 hours, while a “bronze contract” promises an 85% fill rate within two days. In other settings, a supplier segments its customers based on the delivery channel or the price they pay; the supplier recognizes some customers as deserving higher priority over other customers. In other cases, a supplier provides price discounts for delivery flexibility, and then allows a customer to choose the delivery time when placing an order.

A common approach to such scenarios is to categorize the customers into a finite number of demand classes. Customers within a demand class receive the same level of service. The inventory challenge is then to determine how to meet the service-level expectations for each demand class with the least amount of inventory.

In this paper, we consider a single-item inventory system with stochastic demand and multiple demand

classes. The key assumptions are Poisson demand, a deterministic lead time, a continuous-review (Q, R) replenishment policy, and demand backordering. As is common in the literature, we assume a critical-level policy for rationing the inventory across the demand classes.

We have organized this paper into eight sections. In the remainder of this section, we discuss the relevant literature. In the following section, we present our assumptions and a general framework to describe how we manage the inventory with a stationary critical-level policy. In particular, we introduce our allocation policy for clearing backorders. In §3, we show how to map this inventory system into a serial inventory system. In §4, we use this mapping to develop a model for cost evaluation and optimization under the assumptions of Poisson demand, a deterministic replenishment lead time, and a continuous-review (Q, R) policy with rationing. In §5, we pose the service-level problem (SLP), in which we find the critical-level policy that meets a specified fill-rate target for each demand class with the least inventory. Furthermore, we provide a heuristic solution approach for the SLP. We provide some justification for our allocation policy in §6. In §7, we

Table 1 Inventory Literature for Single-Product, Multiple Demand Classes

	Periodic-review, lost sales	Periodic-review, backorders	Continuous-review, lost sales	Continuous-review, backorders
Two demand classes	Evans (1968)	Kaplan (1969) Frank et al. (2003)	Melchiors et al. (2000)	Nahmias and Demmy (1981) Moon and Kang (1998) Dekker et al. (1998) Deshpande et al. (2003) Fadiloglu and Bulut (2005)
N demand classes	Veinott (1965) Topkis (1968)	Atkins and Katircioglu (1996)	Melchiors (2001) Dekker et al. (2002)	Deshpande and Cohen (2005)

provide a numerical experiment both to compare our proposed heuristic with the optimal solution and to show the value from rationing. We show with both bounds and a numerical experiment that this heuristic is quite robust and near optimal. In the final section, we discuss possible extensions and directions for future research. We show how to extend the model to permit service times, whereby different demand classes have different service times by which their demand is to be met. Finally, we describe how to use the single-item inventory system to characterize the inventories and backorders in a multi-echelon distribution system.

Kleijn and Dekker (1998) give an overview of inventory systems with multiple demand classes and provide examples of managing inventory with multiple demand classes, ranging from airline service companies to petrochemical companies. In Table 1, we provide a high-level categorization of the literature. Like much of the stochastic demand inventory literature, we can categorize the research by the control policy, periodic or continuous review, and by the treatment of shortages, lost sales or backorders. In addition, some of the key developments are restricted to or primarily focused on two demand classes, whereas other work is not.

Veinott (1965) analyzes an inventory model with several demand classes for a single product. He proposes to use critical inventory levels to ration the on-hand inventory among demand classes. Topkis (1968) subsequently analyzes the proposed critical-level policy for a periodic-review single-product inventory model with multiple demand classes.

Kaplan (1969) and Evans (1968) study periodic-review models with only two demand classes, similar to Topkis (1968). Recently, Atkins and Katircioglu (1996) and Frank et al. (2003) analyze periodic-review inventory systems with multiple stochastic demand classes. Atkins and Katircioglu (1996) require an associated service level for each demand class, which had not been analyzed in the previous literature. However, their model allows negative inventory allocations that are hard to explain and implement. Frank et al. (2003) apply rationing to avoid incurring high-fixed-ordering

costs rather than saving inventory for high-priority demand.

Nahmias and Demmy (1981) study a continuous-review inventory policy with two demand classes. They assume a (Q, R) inventory replenishment policy, a critical-level policy, and at most one outstanding order at any time. This last assumption implies that whenever a reorder quantity is received, the inventory level and inventory position become identical. This allows them to calculate approximate expressions for expected backorders for both demand classes. Moon and Kang (1998) later extend this model to account for compound Poisson demand processes.

Deshpande et al. (2003) analyze the same (Q, R) inventory rationing model with two demand classes as in Nahmias and Demmy (1981), but without the restriction on the number of outstanding orders. They introduce the threshold clearing mechanism to fill backorders, which permits them to derive expressions for the expected number of backorders for both classes. Based on these expressions, they develop algorithms to calculate the optimal ordering and rationing parameters. They demonstrate numerically the effectiveness of their model, by comparison to a priority-based backlog clearing mechanism, where high-priority backorders are filled before low-priority backorders.

Deshpande and Cohen (2005) extend the analysis in Deshpande et al. (2003) to multiple demand classes under the assumption of the latter paper's threshold clearing policy. They derive expressions for computing the performance measures and state a series of structural results on these performance measures.

Fadiloglu and Bulut (2005) study the inventory rationing problem with two demand classes in Deshpande et al. (2003) using an embedded Markov chain approach. They assume a one-for-one inventory replenishment policy and they clear backorders using a priority clearing mechanism in which they clear class-2 backorders only after restoring the entire reserve stock for class 1. They derive a recursive procedure to calculate the transition probabilities of the corresponding Markov chain; based on this Markov chain, they develop an algorithm for finding the steady-state distribution of the inventory level in the system.

Melchiors et al. (2000) also analyze a (Q, R) inventory model with two demand classes. Unlike Nahmias and Demmy (1981) and Deshpande et al. (2003), they consider a lost sales environment so that demands from the low-priority class are rejected whenever the inventory level drops to the critical level. Melchiors (2001) extends the model in Melchiors et al. (2000) to multiple Poisson demand classes with stochastic replenishment lead times. Moreover, he considers a nonstationary critical-level policy that provides a benchmark to evaluate the stationary critical-level policy employed by Nahmias and Demmy (1981), Melchiors et al. (2000), and Deshpande et al. (2003).

Dekker et al. (1998) study an inventory model with two demand classes and a one-for-one replenishment policy. The model is similar to the one in Nahmias and Demmy (1981). They assume Poisson demand processes, a deterministic replenishment lead time, backordering of unfilled demands, and a critical-level policy to ration the inventory. Dekker et al. (1998) explore how best to handle and allocate incoming replenishment orders, which remains an open question in the literature.

Dekker et al. (2002) extend the model in Dekker et al. (1998) to multiple demand classes with stochastic replenishment lead times by switching to a lost-sales environment rather than by allowing backorders. They assume a one-for-one replenishment policy and a critical-level policy to ration inventory among demand classes. In a lost-sales environment, each incoming replenishment order simply replenishes the inventory. They develop efficient numerical solution methods to calculate the optimal base-stock level and critical levels with or without service-level constraints.

Ha (1997a, b) considers a make-to-stock production system with a single production facility and multiple demand classes for the end product. He assumes exponentially distributed production time, Poisson demand for each demand class, and either lost sales or backorders. He shows that a stationary critical-level policy is optimal. de Véricourt et al. (2000, 2002) consider the multiple-demand class extension of the two-demand class study in Ha (1997a). They characterize the optimal policy for the backorders case with zero setup costs and exponential lead times.

2. General Framework

Our work is most closely related to that of Nahmias and Demmy (1981) and Deshpande et al. (2003). However, whereas their work considers two demand classes, we have no restriction on the number of demand classes. We also develop the model in what we believe is a more transparent and natural way. Indeed, as will be seen, this allows us to extend the

model to permit service times and to analyze a multi-echelon system with multiple demand classes.

We consider a facility that carries inventory for a single product to serve N customer classes. We differentiate customer classes based on their relative service-level requirements or shortage costs. For our analysis, we require the following standard inventory assumptions:

- (i) We have a fixed replenishment lead time $L > 0$;
- (ii) The demand from class i , D_i , $i \in \{1, N\}$, follows a stationary Poisson process with rate λ_i that is independent of the demand from the other demand classes;
- (iii) We replenish inventory with a continuous-review (Q, R) policy; and
- (iv) We backorder any demand that is not immediately met from on-hand inventory.

In addition to these assumptions, we need to describe how we ration inventory across the demand classes. We number the demand classes according to their relative priority, where class 1 has the highest priority. As suggested by Veinott (1965), we use a critical-level policy given by $\mathbf{c} = \{c_1, c_2, \dots, c_{N-1} \mid c_i \in \mathbb{Z}^+ \cup \{0\} \text{ and } c_{i-1} \leq c_i\}$. We stop serving demand class $i + 1$ once the on-hand inventory reaches or falls below the class- i critical stock level c_i ; by assumption, we then backorder all demand for class $i + 1$ until the on-hand inventory is raised above c_i . For class 1, we fill class-1 demand until the on-hand inventory is completely depleted, at which point we backorder any subsequent class-1 demand.

We also need an assumption on how we allocate the inventory replenishment when it is received. The primary issue is how to allocate the replenishment between backorders of different classes and between filling backorders versus restoring the inventory for higher-priority demand classes. To illustrate the challenge, consider a three-class system with order quantity $Q = 4$, and critical levels given by $c_1 = 2$, $c_2 = 3$, $R = 5$. Suppose that just before the order quantity arrives, we have on-hand inventory of one unit, no backorders for class 1, two backorders for class 2, and three backorders for class 3. We need to decide how to allocate the replenishment quantity, as it is insufficient to clear all of the backorders and restore the on-hand inventory above the class-2 critical level $c_2 = 3$. For instance, should the four units be used exclusively to clear backorders, and if so, which of the five backorders should be filled? Should we fill a class-2 backorder before a class-3 backorder, even if the class-3 backorder is older? Alternatively, we might want to hold some units in anticipation of class-1 demand. For instance, if we regard class-1 customers as extremely sensitive to service relative to the other classes, we might want to hold one unit to raise the on-hand inventory to the critical level for class 1 and then use

the remaining three units to clear backorders. Another option is to “ignore” class 3 and use the four units to clear the class-2 backorders and rebuild the on-hand inventory to the class-2 critical level.

As one can see from this limited discussion, there are many possibilities to consider, and clearly the best approach will depend on the context, e.g., the relevant performance metrics, the service constraints, and any policy considerations. In this paper, we propose an allocation mechanism that attempts to strike a balance between the competing trade-offs. To state this mechanism, we need to define the shortfall at time t for each demand class i , $i \in \{1, N - 1\}$ as follows:

$$SF_i(t) = [c_i - IOH]^+ + \sum_{j=1}^i B_{j,j}(t), \quad (1)$$

where IOH is the inventory on-hand at time t and $B_{j,j}(t)$ denotes the backorders for demand class j at time t .¹ The shortfall for class i represents the amount needed to clear all backorders and to restore the on-hand inventory to the critical level for class i . Thus, for the example stated earlier, the shortfalls are $SF_1 = 1$, $SF_2 = 4$.

For our allocation process, we start with class N and decide how to divide the order quantity between backorders for class N and the shortfall for class $N - 1$. We then need to divide the amount allocated to the shortfall for class $N - 1$ between the backorders for class $N - 1$ and the shortfall for class $N - 2$. This process continues until we reach a demand class with no shortfall or we have no more stock to allocate. Between any two adjacent classes, we do not distinguish between a backorder for the lower-priority class and a shortfall for the higher-priority class. We fill these backorders or shortfalls in the order of occurrence, oldest first. In effect, we combine the backorders for the lower-priority class with the shortfalls for the higher-priority class, and we fill these according to a first-come-first-served (FCFS) discipline.

More formally, suppose that a replenishment quantity Q arrives at time t and we have on-hand inventory IOH and backorders $B_{i,i}(t)$, $i \in \{1, N\}$. We determine the shortfall for each class from Equation (1), and we decide the allocation as follows:

Step 1. Set $k := N$; $Q_k := Q$

Step 2. If $Q_k \geq B_{k,k}(t) + SF_{k-1}(t)$, go to Step 3; otherwise go to Step 4.

Step 3. The replenishment quantity is sufficient to fill all class- k backorders, as well as the shortfall at class $k - 1$. Thus, this replenishment clears all backorders and restores the on-hand inventory to the critical level for class $k - 1$. Stop.

Step 4. The replenishment quantity is not sufficient to fill both the class- k backorders and the shortfall at class $k - 1$. By assumption, we combine the backorders for class k with the shortfall for class $k - 1$, and we fill these in the order of occurrence, with no differentiation between a class- k backorder and a class $k - 1$ shortfall. Operationally, we would create a list of all demands from classes i , $i \in \{1, k\}$ subsequent to the time epoch at which $IOH = c_{k-1}$, and then allocate the order quantity to the first Q_k occurrences on this list. Let BF_k denote the number of class- k backorders that are filled. Proceed to Step 5.

Step 5. Stop if $k = 2$. Else, set $Q_{k-1} := Q_k - BF_k$ and $k := k - 1$. Go to Step 2.

For instance, in the example, the allocation proceeds as follows:

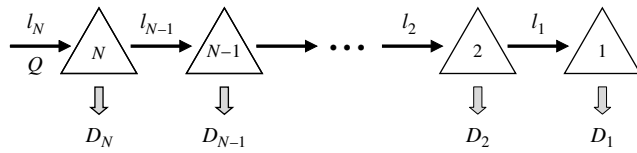
Step 1. $k = 3$, $B_{3,3} = 3$, $SF_2 = 4$, $Q_3 = 4$. We have four units to allocate between class-3 backorders and the class-2 shortfall. We allocate the units in the order of the demand occurrences that created the backorders or shortages. Thus, we track the demand sequence once $IOH = c_2 = 3$, as all subsequent class-3 demand results in a backorder and all subsequent class-1 or class-2 demand results in a shortfall for class 2. Suppose that the demands subsequent to the event $IOH = c_2 = 3$ occur in the order 3, 3, 2, 1, 2, 3, 2, where each number signifies a demand class; then we allocate two units to fill the two oldest class-3 backorders and allocate the remaining two units to the class-2 shortfall.

Step 2. $k = 2$, $B_{2,2} = 2$, $SF_1 = 1$, $Q_2 = 2$. We have two units to allocate between class-2 backorders and the class-1 shortfall. We allocate the units in the order of the demand occurrences that created the backorders or shortages. We need to track the demand sequence once $IOH = c_1 = 2$, as all subsequent class-2 demand results in a backorder and all subsequent class-1 demand results in a shortfall for class 1. From the previous sequence of demands, we infer that the relevant demand order is 1, 2, 2, as the first class-2 demand is served from the protected on-hand inventory for class 2 and we can ignore any class-3 demand occurrences. Thus, we allocate one unit to cover the shortfall for class 1 and we use the other unit to fill the oldest class-2 backorder.

This completes the allocation and we end with $IOH = 2$, $B_{2,2} = 1$, and $B_{3,3} = 1$. We use three units to clear backorders, the two oldest from class 3 and the oldest from class 2, and we use one unit to rebuild the on-hand inventory to the critical level for class 1. In this fashion, we try to balance the objective of clearing backorders of the lower priority classes as fast as we can, with the objective of protecting the higher priority classes from future backorders. Nevertheless, if the order of the demand occurrences were different, we could have a different result.

¹ We drop the time argument for the inventory on-hand, IOH , as it will always be clear from the context; the rationale for the repeated subscript for the backorders will be clearer later, as we develop the model with terminology from the serial stage inventory system.

Figure 1 Serial-Inventory System with Demand at Each Installation



In the next section, we describe how to map the demand class system into a serial stage inventory system. This will facilitate the presentation of the model and its analysis in §4. In §5, after presenting the model, we provide a more detailed justification of this allocation mechanism.

3. The Mapping to a Serial Inventory System

The purpose of this section is to observe the equivalence between the single-product inventory system with N demand classes and a single-product inventory system with N serial stages.

We visualize the demand-class system (DCS) operating as a serial-stage system (SSS), as shown in Figure 1. In effect, we convert the DCS into an SSS by physically separating the on-hand inventory of the DCS into N distinct stockpiles, one for each demand class. We assume that the SSS operates as follows:

(i) Each stockpile or stage i , $i \in \{1, N - 1\}$, operates with a one-for-one continuous review base-stock policy with nonnegative base-stock level s_i , and is replenished by its upstream stage $i + 1$ with replenishment lead time $l_i = 0$. We define $s_i = c_i - c_{i-1}$ with $c_0 = 0$ and interpret the base-stock level to be the *reserve stock* for class i in the DCS, as it represents the stock that is available to serve class- i demand and to replenish the reserve stocks of any higher-priority demand classes.

(ii) Stage N is replenished from an outside supplier with lead time $L > 0$. Stage N uses a continuous-review reorder-point, reorder-quantity policy with its reorder quantity equal to Q and its reorder point equal to $s_N = R - c_{N-1}$, for which we require no assumptions about its sign. We observe that $R = \sum_{i=1}^N s_i$.

(iii) Demand at stage 1 follows a stationary Poisson process with rate λ_1 . Each stage i , $i \in \{2, N\}$, is subject to internal demand from stage $i - 1$, as well as external demand; the external demand at stage i follows a stationary Poisson process with rate λ_i . The N external demand processes are independent.

(iv) At each stage, we backorder all internal and external demand that cannot be met from on-hand inventory.

To establish the equivalence between the DCS and the SSS, we will show that the two systems behave the same (i) when a demand occurs; (ii) when each system places an order on its outside supplier; and

(iii) when each system receives the order from the outside supplier.

(i) *When a demand occurs.* Let IOH represent the on-hand inventory in either the SSS or DCS. If $IOH = 0$, then in both systems we backorder the demand from any class and the on-hand inventory remains at zero.

Consider the DCS and suppose that $IOH > 0$ and $c_{j-1} < IOH \leq c_j$ for some $j \in \{1, N\}$. If the next demand were from class i , $i \in \{1, j\}$, it is served and the on-hand inventory level IOH is reduced by one; if the next demand were from class i , $i \in \{j + 1, N\}$, then it is backordered.

Now consider the SSS with $IOH > 0$ and $c_{j-1} < IOH \leq c_j$. The on-hand inventory at each stage i , $i \in \{1, j - 1\}$ equals its base stock s_i . For each stage i , $i \in \{j + 1, N\}$, there is no on-hand inventory, while stage j has on-hand inventory equal to $IOH - c_{j-1}$. If there was a demand for stage $i \in \{1, j\}$, the serial system fills the demand and the on-hand inventory level IOH is reduced by one; that is, the on-hand inventory at each stage i , $i \in \{1, j - 1\}$ remains at its base stock s_i , and the on-hand inventory at stage j is depleted by one. However, if there was a demand for stage i , $i \in \{j + 1, N\}$, this demand is not filled but is backordered by stage i . The on-hand inventory IOH does not change. Thus, the behavior is the same.

(ii) *When each system places an order on its outside supplier.* In the DCS, we place an order of size Q when the inventory position reaches the reorder point R , where the inventory position is the on-hand inventory, plus the on-order inventory, minus any backorders.

In the SSS, stage N orders from an external supplier when its inventory position reaches a reorder point equal to $s_N = R - c_{N-1}$. We note that the inventory position for each downstream stage $i \in \{1, N - 1\}$ is always $s_i = c_i - c_{i-1}$, due to the one-for-one replenishment policy. Thus, the SSS orders from its external supplier when the system inventory position is $\sum_{i=1}^N s_i = R$. Thus, the two systems behave the same.

(iii) *When each system receives the order from the outside supplier.* Finally, we need to establish that both systems clear the backorders in identical fashion when the replenishment arrives. We state the allocation process for the DCS in §2. To describe the equivalent allocation mechanism for the SSS, we need to define $B_{i+1,i}(t)$, $i \in \{1, N - 1\}$ to be the internal backorders at time t for stage $i + 1$, which represents the number of replenishment requests from stage i that have yet to be filled by stage $i + 1$. From our assumptions on how the SSS operates, we can show that

$$B_{i+1,i}(t) = [c_i - IOH]^+ + \sum_{j=1}^i B_{j,i}(t) = SF_i(t)$$

Thus, we can equate the internal backorders in the SSS to the shortfalls for the DCS. We now assume

that the allocation process for the SSS is as specified for the DCS, but we replace each occurrence of the shortfall SF_{k-1} with the equivalent internal backorder $B_{k,k-1}$. Thus, the two systems behave the same.

This completes the discussion equating the DCS to an SSS. We find this mapping to be helpful in visualizing the operations of the DCS, and in developing an evaluation model of its performance, as described next.

4. Model for N -Demand-Class Inventory System

In this section, we develop a model for evaluating the performance of an N -demand-class inventory system, based on the mapping to a serial system from the prior section. We build this model using the terminology of the SSS, and draw upon the framework in Graves (1985).

We define additional notation to analyze the inventory dynamics in the serial-inventory system, where $i \in \{1, N\}$:

$IL_i(t)$ = inventory level at time t at stage i ;

$IP_i(t)$ = inventory position (inventory level plus inventory on order) at time t at stage i ;

$B_i(t)$ = number of internal and external backorders at time t at stage i ; and

$D_i(s, t)$ = external demand for stage i over interval $(s, t]$.

We can characterize how the inventory level at each stage evolves over time using the following equations for the inventory dynamics for the SSS, where l_i is the replenishment lead time for stage i , and $i \in \{1, N\}$:

$$IL_i(t + l_i) = IP_i(t) - \sum_{j=1}^i D_j(t, t + l_i) - B_{i+1,i}(t), \quad (2)$$

$$B_i(t) = [-IL_i(t)]^+, \quad (3)$$

$$B_i(t) = B_{i,i}(t) + B_{i,i-1}(t), \quad (4)$$

$$B_{1,0}(t) = 0, \quad B_{N+1,N}(t) = 0. \quad (5)$$

The explanation for (2) parallels that in Graves (1985): at time t , the outstanding orders for stage i are either in-process to stage i or are backordered at the immediate upstream stage $i + 1$. By the definition of the lead time, all items that were in-process at time t arrive at stage i before time $t + l_i$, whereas none of the backorders at stage $i + 1$ at time t can arrive to stage i by time $t + l_i$. Furthermore, stage i is subject to demand from its own external demand process, plus that for all downstream stages due to the one-for-one replenishment policy. Any demand during the time interval $(t, t + l_i]$ reduces its inventory level and cannot be replenished by time $t + l_i$. Hence, the inventory

level at time $t + l_i$ at stage i equals its inventory position at time t net of its outstanding orders, namely, the backorders at time t and all demand during the time interval $(t, t + l_i]$.

In Equation (3), we state the backorders to be the negative part of the inventory level. In Equation (4), we decompose the backorders at stage i into backorders created by the external demand at stage i and backorders from replenishment requests from the immediate downstream stage. We stipulate boundary conditions on the model in Equation (5), namely, stage 1 serves no downstream stages and the outside supplier for stage N is reliable and meets any request within its lead time $l_N = L$.

In the context of the DCS, the replenishment lead time $l_i = 0$ for stages i , $i \in \{1, N - 1\}$. Furthermore, due to the continuous-review one-for-one replenishment policy at stages i , $i \in \{1, N - 1\}$, the inventory position for each stage always equals its base-stock level s_i . For stage N , its lead time is positive, $l_N = L$. The steady-state inventory position for stage N is uniformly distributed on the range $[s_N + 1, s_N + Q]$, given the assumption of a (Q, R) system with these parameters (Zipkin 2000, p. 193). We use these observations to rewrite the steady-state form for Equations (2)–(5):

$$IL_N = IP_N - \sum_{i=1}^N D_i^L, \quad (6)$$

$$IL_i = s_i - B_{i+1,i} \quad \text{for } i \in \{1, N - 1\}, \quad (7)$$

$$B_i = [-IL_i]^+, \quad (8)$$

$$B_i = B_{i,i} + B_{i,i-1}, \quad (9)$$

$$B_{1,0} = 0, \quad (10)$$

where D_i^L is the random variable for the external demand at stage i over an interval of length L ; thus, it represents a Poisson random variable with mean $\lambda_i L$.

We need to establish one more property before we can use Equations (6)–(10) to determine the steady-state distribution of the inventory level at each stage. We intend to use (6) or (7) to find the distribution of the inventory level, and then (8) to get the distribution of the total backorders B_i at a stage. We then need to find the distribution of $B_{i,i-1}$, the internal backorders at stage i due to downstream demand. To do this, we contend that the probability distribution of $B_{i,i-1}$, conditioned on a realization for B_i , is a binomial. In particular, we have for $j \in \{0, n\}$ that

$$\Pr[B_{i,i-1} = j \mid B_i = n] = \binom{n}{j} p_i^j (1 - p_i)^{n-j}$$

where $p_i = \frac{\sum_{j=1}^{i-1} \lambda_j}{\sum_{j=1}^i \lambda_j}$. (11)

As an explanation, we note that once stage i stocks out, backorders occur randomly according to the rates

for the Poisson demand processes. The backorders due to external demand at stage i occur at rate λ_i ; backorders due to internal demand from stage $i - 1$ occur at rate $\sum_{j=1}^{i-1} \lambda_j$. Thus, if n backorders occur, the number of backorders due to internal demand is a binomial random variable with parameters (n, p_i) . Furthermore, the allocation scheme for filling backorders, described in the prior section, preserves this random distribution of backorders, as it fills backorders in the order of their occurrence. As a consequence, at any time t , if stage i has positive backorders, then $B_i(t) = \sum_{j=1}^i D_j(s, t)$ for some value of $s < t$. Due to the memoryless property and independence of the Poisson demand processes, the conditional distribution of $B_{i,i-1}$ is binomial.

We can now determine the steady-state distribution of the inventory levels, given the policy parameters (Q, R) and $\mathbf{s} = (s_1, \dots, s_N)$, where $s_N = R - \sum_{i=1}^{N-1} s_i$. The procedure starts from the most upstream stage N and moves iteratively to each downstream stage, as follows:

Step 1. Determine the steady-state distribution of IL_N . We obtain the distribution of IL_N from Equation (6) by convolving the distribution of IP_N with that for $\sum_{i=1}^N D_i^L$. The former is a uniform discrete random variable on the interval $[s_N + 1, s_N + Q]$; the latter is a Poisson random variable with mean $L \sum_{i=1}^N \lambda_i$.

Set $i = N$.

Step 2. Obtain the steady-state distribution of $B_i = [-IL_i]^+$, backorders at stage i .

Step 3. Determine the steady-state distribution of $B_{i,i-1}$. We use the distribution for B_i with (11) to get the unconditioned distribution for $B_{i,i-1}$.

Step 4. Set $i := i - 1$. Determine the steady-state distribution of IL_i from (7).

Step 5. Stop if $i = 1$. Otherwise go to Step 2.

With the steady-state distribution of the inventory level at each stage, we can compute relevant performance measures, such as the expected on-hand inventory, the expected backorders, and the fill rate for each demand class. We can then pose an optimization problem to find the best choice for the control parameters, namely, the reorder point R , reorder quantity Q , and the reserve stock levels $\{s_i: i \in \{1, N - 1\}\}$ (or equivalently, the critical stock levels $\{c_i = \sum_{j=1}^i s_j, i \in \{1, N - 1\}\}$). In the next section, we illustrate one such optimization, in which we minimize the expected on-hand inventory subject to constraints on the fill rates for each demand class.

5. Service-Level Problem

There are many ways to look at the trade-off between holding inventory and achieving a high level of customer service. We consider one problem variant, in which we minimize the amount of inventory needed

to satisfy a given fill-rate target for each demand class. In effect, we define the demand classes by their fill-rate targets; we would cluster customers into the demand classes according to their service expectations, with demand class 1 corresponding to the highest level of service and so on. We formulate this service level problem (SLP) for a DCS as follows:

$$\begin{aligned} \text{SLP} \quad & \text{Min } z = \sum_{i=1}^N E[IL_i]^+ \\ & \text{s.t. } \text{Fillrate}_i \geq \beta_i \quad \text{for } i \in \{1, N\}, \\ & s_i \geq 0, \quad \text{integer for } i \in \{1, N - 1\}, \\ & s_N, \quad \text{integer,} \end{aligned}$$

where

$$\begin{aligned} \text{Fillrate}_i &= \begin{cases} \Pr(IL_i > 0) & \text{if } s_i > 0 \\ \text{Fillrate}_{i+1} & \text{if } s_i = 0 \end{cases} \quad \text{for } i \in \{1, N - 1\}, \\ \text{Fillrate}_N &= \Pr(IL_N > 0). \end{aligned}$$

The objective is to minimize the expected on-hand inventory, which is the positive part of the inventory level. The reserve stocks s_i are the decision variables, from which we can find both the critical levels c_i and the reorder point R . To simplify the presentation, we assume that the order quantity Q is not a decision variable, but has been prespecified.

The constraints assure that we meet a fill-rate target β_i for each demand class i . The computation of the fill rate depends on the reserve stock level of the class. If the reserve stock is positive, then for Poisson demand the fill rate equals the probability that the inventory level is positive. When the reserve stock for demand class i is zero, there is no distinction in order fulfillment between a demand from class i and a demand from class $i + 1$; in this case, the fill rate for demand class i is the same as that for demand class $i + 1$.

In formulating the SLP, we expect (although we do not require) that the higher-priority demand classes have larger fill-rate targets; that is, we expect $\beta_1 \geq \beta_2 \geq \dots \geq \beta_N$. Indeed, the structure of the critical-level policy guarantees that demand class i has a fill rate no worse than that for demand class $i + 1$.

From model (6)–(10), we see that the inventory level at demand class i , IL_i , depends on its reserve stock and that for lower-ranked demand classes; that is, IL_i is a function of (s_i, \dots, s_N) . In the following, we will at times use the notation $IL_i(s_i, \dots, s_N)$ to make this dependence explicit.

5.1. Solution Procedure for the Service-Level Problem

In this section we state a sequential solution method, the single-pass-algorithm (SPA), which provides us with a good feasible solution for the SLP. We then

establish a bound on the gap between the SPA solution and the optimal solution to the SLP.

For each demand class i , SPA uses model (6)–(10) to find the minimum value for its reserve stock that satisfies its fill-rate target, given the previously determined reserve stocks for demand classes $i + 1, \dots, N$. We state the algorithm as follows:

Step 1. Find reserve stock and fill rate for stage N :

$$\hat{s}_N = \min\{s: \Pr(IL_N(s) > 0) \geq \beta_N\};$$

and $Fillrate_N = \Pr(IL_N(\hat{s}_N) > 0)$; let $i := N - 1$.

Step 2. Find reserve stock and fill rate for stage i :

- (a) if $Fillrate_{i+1} \geq \beta_i$: $\hat{s}_i = 0$; $Fillrate_i = Fillrate_{i+1}$;
- (b) if $Fillrate_{i+1} < \beta_i$:

$$\hat{s}_i = \min\{s: \Pr(IL_i(s, \hat{s}_{i+1}, \dots, \hat{s}_N) > 0) \geq \beta_i\};$$

and $Fillrate_i = \Pr(IL_i(\hat{s}_i, \hat{s}_{i+1}, \dots, \hat{s}_N) > 0)$.

Step 3. Stop if $i := 1$. Otherwise, let $i := i - 1$ and repeat Step 2.

The SPA yields a feasible solution for the SLP by construction: at each iteration, it sets the reserve stock for a demand class to satisfy the fill-rate constraint for this demand class. However, there is no guarantee that the solution is optimal; later in this section, we provide an example that illustrates this.

We contend that the solution from the SPA should be quite good. To develop this argument, it will be helpful to rewrite the objective function of the SLP as

$$z = \sum_{i=1}^N E[IL_i]^+ = \sum_{i=1}^N (s_i + E[B_{i,i}]) + \frac{Q+1}{2} - L \sum_{i=1}^N \lambda_i. \quad (12)$$

We obtain this expression from substituting (6), (7), and (9) into $E[IL_i]^+ = E[IL_i] + E[B_i]$, with the observation that $E[IP_N] = s_N + (Q + 1)/2$. Thus, the objective function consists of the sum of the reserve stocks and the sum of the external backorders, plus a constant K :

$$z = \sum_{i=1}^N s_i + \sum_{i=1}^N E[B_{i,i}] + K. \quad (13)$$

We will develop a bound on z by finding a lower bound on the sum of the reserve stocks, and then a lower bound on the sum of the external backorders.

To explore the quality of the SPA solution, we state three propositions; the proofs for these propositions can be found in the online appendix (provided in the e-companion).² We first establish that moving one unit of reserve stock from class j to class $j - 1$ cannot decrease the inventory level at any of the higher-priority classes, but can result in more backorders.

PROPOSITION 1. Consider two stocking policies \mathbf{s}^1 and \mathbf{s}^2 , where for some j , $s_j^2 = s_j^1 - 1$, $s_{j-1}^2 = s_{j-1}^1 + 1$, and $s_i^2 = s_i^1$

$\forall i \neq j, j - 1$. We have that

- (i) $IL_i(s_i^2, \dots, s_N^2) \geq IL_i(s_i^1, \dots, s_N^1)$ for $i \in \{1, j - 1\}$ and
- (ii) $\sum_{i=1}^j B_{i,i}(s_i^2, \dots, s_N^2) \geq \sum_{i=1}^j B_{i,i}(s_i^1, \dots, s_N^1)$.

We now use this proposition to establish bounds on the optimal solution. The next proposition states that the solution for the SPA provides a series of lower bounds on successive sums of the reserve stocks.

PROPOSITION 2. For all feasible solutions \mathbf{s} for the SLP, we have $\sum_{i=j}^N s_i \geq \sum_{i=j}^N \hat{s}_i$ for all j , where $\hat{\mathbf{s}}$ is the solution found by the SPA.

As a special case of this proposition, we see that $\sum_{i=1}^N \hat{s}_i = \hat{R}$ is a lower bound on the sum of the reserve stocks in the objective function of the SLP. Proposition 3 provides an ordering of the objective function values for stocking policies with equal reorder points, i.e., with $\sum_{i=1}^N s_i = R$ for some constant R .

PROPOSITION 3. Consider two stocking policies \mathbf{s}^1 and \mathbf{s}^2 with $\sum_{i=1}^j s_i^1 \leq \sum_{i=1}^j s_i^2 \forall j \in \{1, N - 1\}$ and $\sum_{i=1}^N s_i^1 = \sum_{i=1}^N s_i^2$. Then, we have $z(\mathbf{s}^1) \leq z(\mathbf{s}^2)$.

From this proposition, we have a lower bound on the optimal objective function value of the SLP:

$$\hat{R} + \sum_{i=1}^N E[B_{i,i}(s_i = 0, s_{i+1} = 0, \dots, s_{N-1} = 0, s_N = \hat{R})] + K. \quad (14)$$

Based on these propositions, we conjecture that the solution for the SPA should be near optimal in most settings. We see from the lower bound that any improvement to the SPA solution must come by means of a reduction in backorders. As most settings have high-service expectations, the fill-rate targets are such that any feasible solution will generate, at most, a modest amount of backorders. As a consequence, we expect there to be minimal opportunity to improve upon the solution given by the SPA. We explore this conjecture in §7 with a computational experiment.

EXAMPLE. The purpose of this example is to provide some insight into why the SPA solution need not be optimal, yet is likely to be close to optimal.

We assume three demand classes with the Poisson demand rates $\lambda_1 = 8$ units/year, $\lambda_2 = 12$ units/year, and $\lambda_3 = 16$ units/year. The replenishment lead time is $L = 1/4$ year (three months), and the reorder quantity is $Q = 1$. The fill-rate targets are $\beta_1 = 0.99$, $\beta_2 = 0.94$, and $\beta_3 = 0.87$.

When we apply the SPA to this problem, we get $\hat{s}_1 = 2$, $\hat{s}_2 = 1$, $\hat{s}_3 = 12$ with $z = 7.09$. This translates into the critical-level policy $\hat{c}_1 = 2$, $\hat{c}_2 = 3$, and $\hat{R} = 15$; we reorder when the inventory position reaches 15, and we stop serving demand classes 3 and 2 once the on-hand inventory drops to 3 and to 2, respectively.

² An electronic companion to this paper is available as part of the online version that can be found at <http://mansci.journal.informs.org/>.

We can use (12) to break the objective value into its constituent parts:

$$z(\hat{s}_1, \hat{s}_2, \hat{s}_3) = \sum_{i=1}^3 \hat{s}_i + \sum_{i=1}^3 E[B_{i,i}] + \frac{Q+1}{2} - L \sum_{i=1}^N \lambda_i \\ = 15 + 0.09 + 1 - 9 = 7.09.$$

As the expected backorders are quite small, we know from Proposition 3 that this solution must be very close to optimal. Indeed, the lower bound from Proposition 3 is 7.02.

The optimal solution (found by exhaustive search) is $s_1 = 1$, $s_2 = 0$, $s_3 = 14$ with $z = 7.03$. The only difference between the optimal solution and the SPA solution is that we move one unit of reserve stock from both demand class 1 and demand class 2 to demand class 3; these moves reduce the backorders at demand class 3 without jeopardizing the fill-rate constraint for demand classes 1 and 2. These moves do reduce the fill rate for demand class 1, but it still is above 0.99.

To appreciate the benefit of differentiating the demand classes, suppose that all customers were to get the highest service level, namely, a fill rate of 0.99. Then, we need to set the reorder point $R = 17$ and the expected on-hand inventory (z) is 9.00, which is 28% higher than the optimal solution.

In general, we find that we often can slightly improve the SPA solution by shifting a few units of reserve stock from higher-priority classes to lower-priority classes. The result of these shifts is that we get a slight reduction in the total backorders: we reduce some backorders at the lower-priority demand classes, yet still satisfy the fill-rate target at the higher-priority classes.

6. Justification for the Allocation Process

In this section, we provide some justification for the allocation process. For the DCS, the intent of the allocation process is to balance the trade-offs between filling current backorders and rebuilding the reserve stock so as to avoid future backorders. For each pair of adjacent demand classes, we treat a backorder for a lower-priority class equivalent to a reserve-stock shortfall for the higher-priority class. We see that for the SSS, this process results in allocating the inventory at each stage to the internal and external backorders at the stage in the order of occurrence. In the following, we provide five arguments in support of this allocation process.

First, we regard the determination of an optimal allocation process to be a most challenging proposition. For instance, for the SLP we suspect that the optimal allocation policy would be a dynamic policy that depends on the state of the system. Furthermore,

we expect that the optimal policy would opt to not fill many backorders; as we have defined the problem, the only reason to fill a backorder is to lower the on-hand inventory and reduce the inventory holding costs. As this is likely to be a nonpractical outcome, the SLP seems ill-specified for the purposes of deciding an optimal allocation policy; we need to add something to the problem statement to force us to fill the backorders in some reasonable way. For the sake of argument, one might augment the SLP with the requirement that we fill the backorders and shortfall at each demand class in an FCFS fashion, in which case our rule is optimal. Alternatively, we could redefine the SLP with the requirement that we use some form of a priority clearing mechanism and/or add backorder costs to the objective function; in this case we expect our rule does quite well and is near optimal, as we discuss below.

Second, our FCFS allocation rule is equivalent to the threshold clearing process developed by Deshpande et al. (2003) and Deshpande and Cohen (2005). In the appendix, we develop the correspondence between the sample paths for the two rules. In the online appendix, we show the algebraic equivalence of the steady-state performance measures for a two-class system. Deshpande et al. (2003) conduct an extensive simulation study comparing the performance of the threshold clearing mechanism to the priority clearing mechanism in which class-1 backorders are always cleared before class-2 backorders. They assume a context with a setup cost and inventory holding cost, and a class-dependent backorder cost that is linear in the duration of the backorder. They find virtually no difference between their threshold clearing process and the priority clearing mechanism when there is a high or moderate setup cost; for the case of no setup cost ($Q = 1$), they find that the cost difference increases with the spread between the backorder costs for the two classes and can be as large as 5% or 6%. Nevertheless, the overall conclusion is that the threshold clearing process performs very well.

Third, it is possible for our allocation rule to make a “nonpriority allocation.” That is, we clear an older backorder for a lower-priority class and leave unfilled a more recent backorder for a higher-priority class. Dependent upon the performance metrics for the system, this may or may not be the preferred allocation. Nevertheless, these nonpriority allocations represent a primary difference between our FCFS rule and a priority clearing rule. We show in the online appendix how to compute the probability of a nonpriority allocation for a two-class system with $Q = 1$; we choose this case as it is the most demanding test for the FCFS rule, and results in the highest probability for a nonpriority allocation. We also report in the online appendix a numerical test in which we evaluate the

Table 2 Fill-Rate Comparisons Between FCFS and the Priority Allocation Rules

Number of classes	Number of comparisons	Mean difference	Max difference	Mean percent difference (%)	Max percent difference (%)
$N = 2$	16	0.005	0.013	0.5	1.3
$N = 3$	96	0.010	0.040	1.2	5.5
$N = 4$	204	0.011	0.040	1.3	5.5
$N = 5$	224	0.013	0.037	1.5	4.4

probability for a nonpriority allocation for 209 test cases for the SLP. We find that nonpriority allocations do not occur very frequently. The maximum probability for a nonpriority allocation in our test problems is 0.07. The probability of a nonpriority allocation is higher than 0.05 in only 3% of the test problems and is less than 0.01 in 75% of the test problems. The average value of the probability over the test problems is 0.01. Thus, we find that there is little difference between our allocation rule and a priority clearing rule when $Q = 1$ for these test problems.

The fourth point is that the FCFS allocation process is effectively the same as the virtual allocation mechanism introduced in Graves (1996) for the analysis of a multi-echelon arborescent inventory system. For the virtual allocation mechanism, each stage in the system allocates inventory to its customers, both internal and external, on an FCFS basis, that is, in the order of the demand occurrences. This is the same as we have assumed for the DCS. For a two-echelon system with one central warehouse and multiple retailers, Graves (1996) finds that virtual allocation yields near-optimal inventory requirements in comparison with a lower bound. This provides some additional evidence that FCFS allocation is effective, albeit for a different problem setting.

For the final argument, we compare the FCFS allocation rule to a priority allocation rule on a set of test problems. For each test problem, we set the order quantity $Q = 4$ and the lead time $L = 1/4$ year. The demand rate is $\lambda = 36$ units per year, and we have four possible settings for the demand classes and demand rates:

$$\begin{aligned} &\{N = 2; \lambda_1 = 18, \lambda_2 = 18\}, \\ &\{N = 3; \lambda_1 = 8, \lambda_2 = 12, \lambda_3 = 16\}, \\ &\{N = 4; \lambda_1 = 4, \lambda_2 = 6, \lambda_3 = 10, \lambda_4 = 16\}, \\ &\{N = 5; \lambda_1 = 4, \lambda_2 = 6, \lambda_3 = 8, \lambda_4 = 8, \lambda_5 = 10\}. \end{aligned}$$

For each demand setting, we set the reserve stock for the lowest-priority class as $s_N = 7, 8, 9$, or 10 . Setting $s_N = 7$ results in a 0.52 fill rate for class N , while setting $s_N = 10$ gives a fill rate of 0.83 for class N . We show in the online appendix that the fill rate for class N is the same for both FCFS and priority allocation rules.

For each demand setting, we generate a test problem by enumerating all of the possible reserve stock settings (s_1, \dots, s_N) that satisfy

$$\begin{aligned} s_i \geq 1, \quad i = 1, \dots, N; \quad s_N = 7, 8, 9, 10; \quad \text{and} \\ \sum_{i=1}^N s_i = 11, 12, 13, 14. \end{aligned}$$

This yields 188 test problems. We consider these reserve stock settings as they yield realistic fill rates, ranging from 0.52 to 0.99.

For each test problem, we compare the fill-rate performance of the FCFS allocation rule to that for the priority allocation rule in which we fill all backorders and restore the reserve stock for a higher-priority class before allocating any inventory to backorders for a lower-priority class. We choose this form of the priority rule as we expect it gives the highest fill rates for the high-priority classes and thus will generate the largest discrepancy with the FCFS rule. We only compare the fill rates for classes $1, \dots, N - 1$, as the fill rates are the same for class N .

We use model (6)–(10) to compute the fill rates for each test problem for the FCFS allocation rule. For the priority allocation rule, we employ a ProModel™ simulation to estimate the fill rates. For these tests, we ran the simulation so that the 95% confidence interval for each fill-rate estimate is ± 0.001 .

We report the fill-rate results of the experiment in Table 2. The number of comparisons is the number of test problems multiplied by $N - 1$, as we compare the fill rates for classes $1, 2, \dots, N - 1$ for each test problem. The fill rates for the priority allocation rule are always as great as the fill rate for the FCFS allocation rule for these test problems (see the online appendix). We report the mean difference both in absolute terms and as a percentage, as well as the maximum difference.

For these test problems, we find the priority rule does better but the increase in fill rate is quite small. Under the priority rule, the fill rate for class-1 increases on average by 0.5% for $N = 2$; the fill rates for class-1 and class-2 increase on average by 1.2% for $N = 3$; the fill rates for class-1, class-2, and class-3 increase on average by 1.3% for $N = 4$; and the fill rates for class-1 through class-4 increase on average

Table 3 Expected Backorder Comparisons Between FCFS and the Priority Allocation Rules

Number of classes	Number of comparisons	Mean difference	Max difference	Mean percent difference (%)	Max percent difference (%)
$N = 2$	16	0.057	0.124	17.0	23.7
$N = 3$	96	0.064	0.129	16.7	27.8
$N = 4$	204	0.073	0.129	16.6	27.7
$N = 5$	224	0.082	0.123	23.0	39.4

by 1.5% for $N = 5$. The maximum increase in fill rates ranges from 1.3% to 5.5%.

There is a cost associated with the fill-rate improvement from the priority allocation rule. The priority rule results in a longer duration for backordered demand relative to the FCFS rule. To appreciate this, we report in Table 3 the increase in the expected backorder level from priority allocation, relative to the FCFS rule. The expected backorder level for the priority rule has increased on average by more than 16% in comparison to the FCFS rule. As there are fewer backordered demands for the priority allocation, we can infer from Little’s Law that the length of time to clear a backorder is at least 16% higher on average for the priority allocation rule.

The computational effort to evaluate the priority rule is several orders of magnitude greater than that for the FCFS rule; with an Intel® Pentium® M processor 1.1 GHz, 454 MHz, 504 MB of RAM computer, the time to compute the fill rates for FCFS is instantaneous for each test problem, whereas for the priority rule the time is about 22.5 minutes (10 replications, each at 2.25 minutes) using the ProModel™ simulation.

7. Test of Effectiveness of the SPA

To test the effectiveness of the SPA on the SLP, we compare its solution to the optimal solution on two sets of test problems in two experiments. For the first experiment, we examine the performance of the SPA as we vary the reorder quantity, the lead time, the fill-rate targets, and the demand rates. For the second experiment, we vary the number of demand classes.

For each test problem in the first experiment, there are three demand classes. The reorder quantity takes on one of four values: $Q = 1, 4, 9, \text{ or } 18$. The replenishment lead time from the outside supplier is one of the three values: $L = 1/24$ year, $1/4$ year, $1/2$ year. There are three possible values for the fill-rate target for each of the demand classes: $\beta_1 = 0.9, 0.95, \text{ or } 0.99$; $\beta_2 = 0.8, 0.9, \text{ or } 0.95$; and $\beta_3 = 0.7, 0.8, \text{ or } 0.9$. We only consider combinations with either $\beta_1 > \beta_2 \geq \beta_3$ or $\beta_1 \geq \beta_2 > \beta_3$; thus, we have 20 combinations of fill rates. Finally, we have four possible settings for the demand rates: $\{\lambda_1 = 8, \lambda_2 = 12, \lambda_3 = 16\}$, $\{\lambda_1 = 16, \lambda_2 = 12, \lambda_3 = 8\}$, $\{\lambda_1 = 1, \lambda_2 = 3, \lambda_3 = 8\}$, and $\{\lambda_1 = 4, \lambda_2 = 4, \lambda_3 = 4\}$ units/year.

We specify a test problem in the first experiment by setting the number of demand classes (one candidate), the replenishment lead time (three candidates), the reorder quantity (four candidates), the set of desired fill rates (20 candidates), and the set of demand rates (four candidates). This provides a total of 960 test problems.

For each test problem, we compute the SPA solution $(\hat{s}_1, \hat{s}_2, \hat{s}_3)$ and its cost from Equation (13); the lower bound from Equation (14); and the optimal solution and its cost. We find the optimal solution by a search algorithm. We first compute $z(s_1 = 0, s_2 = 0, s_3 = \hat{s}_1 + \hat{s}_2 + \hat{s}_3 + 1)$, which is a lower bound on the cost for any solution for the SLP with reorder point $R = \hat{s}_1 + \hat{s}_2 + \hat{s}_3 + 1$. In all test problems, we find $z(\hat{s}_1, \hat{s}_2, \hat{s}_3)$ to be less than $z(s_1 = 0, s_2 = 0, s_3 = \hat{s}_1 + \hat{s}_2 + \hat{s}_3 + 1)$. This observation together with the results in Proposition 2 and Proposition 3 guarantees that the reorder point in the optimal solution must be $\hat{s}_1 + \hat{s}_2 + \hat{s}_3$. Next, we find the optimal solution by searching over the integer solutions in the space: $s_3 \geq \hat{s}_3$; $s_2 + s_3 \geq \hat{s}_2 + \hat{s}_3$; $s_1 + s_2 + s_3 = \hat{s}_1 + \hat{s}_2 + \hat{s}_3$; and $s_1, s_2 \geq 0$.

The results of this numerical experiment support our intuition that the SPA is quite effective. The SPA finds the optimal solution in 274 problem instances or 29% of the cases. The cost of the SPA solution is on average 0.57% higher than the cost of the optimal solution and 1.28% higher than the lower bound. The maximum error for the SPA is 3.24%.

In Table 4, we examine how the relative performance of the SPA heuristic changes as we vary the problem parameters. Each cell of the table provides the average cost increase for the SPA solution for all test problems with the single parameter fixed. For instance, in the cell with $L = 1/24$, we report the average performance of the SPA for the 320 test problems with lead time $L = 1/24$. For the fill-rate targets, we have divided the 20 combinations according to the spread between the fill-rate targets for class 1 and class 3 ($\beta_1 - \beta_3$). There are six combinations and 288 test problems with $0.05 \leq \beta_1 - \beta_3 < 0.15$, eight combinations (384 problems) with $0.15 \leq \beta_1 - \beta_3 < 0.25$, and six combinations (288 problems) with $0.25 \leq \beta_1 - \beta_3$.

The performance of the SPA seems quite insensitive to the settings for the reorder quantity and the replenishment lead time. However, the performance seems to depend on the distribution of demand rates

Table 4 Average Increase in the SPA Solution Relative to the Optimal Solution

	$L = 1/24$	$L = 1/4$	$L = 1/2$	
Lead time	0.52%	0.66%	0.54%	
Reorder quantity	$Q = 1$ 0.58%	$Q = 4$ 0.56%	$Q = 9$ 0.58%	$Q = 18$ 0.57%
Demand rates	$\{\lambda = 8, 12, 16\}$ 0.64%	$\{\lambda = 16, 12, 8\}$ 0.46%	$\{\lambda = 1, 3, 8\}$ 0.65%	$\{\lambda = 4, 4, 4\}$ 0.53%
Service target spread	$0.05 \leq \beta_1 - \beta_3 < 0.15$ 0.32%	$0.15 \leq \beta_1 - \beta_3 < 0.25$ 0.56%	$0.25 \leq \beta_1 - \beta_3$ 0.84%	

and the spread in fill-rate targets. The performance improves slightly when there is a higher percentage of demand in the higher-priority demand class (class 1). In addition, the SPA performs best when the spread in service levels is smallest.

For each test problem, we also compute the cost for the optimal inventory policy in which we provide the highest fill rate β_1 for each demand class. Admittedly, this is a suboptimal policy as there is no rationing of inventory between demand classes; nevertheless, we observe this policy in practice as it satisfies the service requirements and is easy to implement. For these test problems, the cost of a no-rationing policy is on average 18% higher than the optimal critical-level policy.

In the second experiment, we specify four test problems, one for each setting for the number of demand classes: $N = 2, 3, 4$, or 5 . We set the replenishment lead time $L = 1/4$ year and the reorder quantity $Q = 4$ for each test problem. In Table 5, we specify the fill-rate targets and the demand rates for each test problem. We compute the SPA solution and cost for each test problem, as well as the optimal solution and its cost. We report the results in Table 5. On this set of test problems, the relative performance of the SPA heuristic improves as the number of demand classes increases. This observation is consistent with our intuition that the SPA performs better when there are smaller differences between the fill-rate targets for consecutive demand classes.

8. Extensions

In this paper, we consider a single-product inventory system with multiple demand classes. We show how to map this system into an equivalent single-product serial-inventory system. We then apply a modeling

framework for multi-echelon divergent systems to obtain a characterization of the steady-state performance of the N -demand-class inventory system for a critical-level policy. To find the best critical-level policy, we pose an optimization problem to minimize the on-hand inventory subject to fill-rate constraints for each demand class. We provide a computationally-efficient approximate procedure for solving this problem, and demonstrate its effectiveness on a set of test problems. In this section, we first show how to incorporate service times into the model and how to use the model to characterize a multi-echelon distribution system with multiple demand classes. We then briefly discuss possible extensions to this research.

8.1. Service Times

In the presentation so far, we assume that the service time for each demand class is zero. That is, customers in each demand class expect their demand to be filled at the time of its occurrence. In many contexts, however, it is common to have nonzero service times, whereby a customer expects demand to be filled within some specified time window. Indeed, this can be the basis for defining demand classes. Demand class 1 might be the customers, who, say, have a 24-hour service time. The other demand classes might have longer service times, say three days for class 2, one week for class 3, and so on. For instance, for service-parts inventory systems, these service times are part of the contract between the customer and the provider of the service parts. Another example is where customers select the time of delivery, as is available from most e-tailers. In this manner, the customer defines (and pays for) a desired service time.

We need to describe how the critical-level policy applies when service times are nonzero. Let w_i be the service time for demand class i . We assume that each

Table 5 Test Problems and Results from the Second Computational Experiment

N	Service targets	Demand rates	SPA solution	Optimal solution	Percent difference (%)
2	$\{\beta = 0.99, 0.8\}$	$\{\lambda = 18, 18\}$	7.627	7.542	1.13
3	$\{\beta = 0.99, 0.9, 0.8\}$	$\{\lambda = 8, 12, 16\}$	6.646	6.583	0.96
4	$\{\beta = 0.99, 0.95, 0.9, 0.8\}$	$\{\lambda = 4, 6, 10, 16\}$	6.644	6.587	0.86
5	$\{\beta = 0.99, 0.95, 0.9, 0.85, 0.8\}$	$\{\lambda = 4, 6, 8, 8, 10\}$	6.628	6.591	0.56

$w_i < L$. We say that a demand from class i that arrives at time t is due at time $t + w_i$. Let IOH represent the on-hand inventory in the system. Suppose that $c_{j-1} < IOH \leq c_j$ for some $j \in \{1, N\}$. Then, if the next demand due was from class i , $i \in \{1, j\}$, it is served and the inventory level IOH is reduced by one; if the next demand due was from class i , $i \in \{j+1, N\}$, then it is backordered.

This policy ignores information about realized demand that is not yet due. Nevertheless, the policy would be relatively easy to implement and does allow for stock rationing so as to protect the service to higher-priority demand classes.

We do assume that when a demand arrives from any demand class, the system inventory position is reduced by one, and a reorder is placed once the system inventory position reaches the reorder point.

We also need to indicate how the allocation mechanism will work. For each demand class, we again define its shortfall at any point in time by Equation (1). The allocation mechanism operates the same as for the zero service-time case. For each pair of adjacent classes, we combine the backorders for the lower-priority class with the shortfalls for the higher-priority class; we then fill these backorders or shortfalls in the order of occurrence, but now we use the due dates for the demand occurrences.

With these assumptions, we can restate the steady-state equations analogous to (6) and (7):

$$IL_N = IP_N - \sum_{i=1}^N D_i(L - w_i), \quad (15)$$

$$IL_i = s_i - B_{i+1,i} \quad \text{for } i = 1, 2, \dots, N-1, \quad (16)$$

where $D_i(\tau)$ is the random variable for the external demand at stage i over an interval of length τ ; thus, it represents a Poisson random variable with mean $\lambda_i\tau$. Equations (8), (9), and (10) are unchanged. We can then use the algorithm from §4 to evaluate these equations to find the steady-state distribution of the inventory levels, given the policy parameters (Q, R) , $(c_1, \dots, c_N - 1)$, and (w_1, \dots, w_N) .

We thus see that model (6)–(10) extends directly to permit nonzero service times, with the same computational requirements. Nevertheless, there is an open question as to how effective is the (myopic) critical-level policy for this extension.

8.2. Multi-Echelon Systems

As a second extension, we describe how one might develop a model of a multi-echelon inventory system with multiple demand classes. For instance, consider a service-parts distribution system in which there is a central warehouse that replenishes several local sites. We assume that each of the local sites is subject to Poisson demand from N classes, operates with a

critical-level policy, and reorders on the central warehouse with an order quantity $Q = 1$. We assume that the central warehouse replenishes its inventory with a one-for-one replenishment from an external supplier with a deterministic lead time, and fills the order from the local sites on an FCFS basis with a deterministic lead time. These assumptions are typical for low-volume, high-value service parts.

We can use model (6)–(10) for each site, but with one modification. When the central warehouse stocks out, replenishment requests from the local sites are delayed. Thus, we need restate Equation (6) for the inventory at local site k as

$$IL_{N,k} = IP_{N,k} - \sum_{i=1}^N D_{i,k}^L - B_{0,k}, \quad (17)$$

where the second subscript refers to the local site, and where $B_{0,k}$ denotes the backorders at the central warehouse that are due to local site k . For Poisson demand, we can use either the exact or approximate model in Graves (1985) to characterize the backorders at the central warehouse as a function of its base-stock level.

Thus, we can model the performance of a multi-echelon system with N -demand classes for Poisson demand, one-for-one replenishment policies, and deterministic lead times. We can use this model to optimize the inventory parameters, namely, the base stock at the central warehouse and the critical levels and reorder point at each of the local sites. One approach would be to do a single-dimension search over possible settings for the base stock at the central warehouse. Given a base stock at the central warehouse, we can characterize the backorders to each of the local sites. We can then use (17) and (7)–(10) to optimize the inventory parameters at each local site, as described in this paper.

8.3. General Demand Process

We assume a Poisson demand process. Model (2)–(5) remains valid for demand processes with independent increments, e.g., compound Poisson demand. However, we would need to revisit the next steps in the model development if demand was from a compound Poisson process. The distribution of the inventory position in Equation (6), IP_N , is no longer uniform, as it will depend on the compounding distribution. Similarly, the conditional distribution of backorders for demand class i due to downstream demand is not binomial, as given by (11). The computation of fill rate at each demand class is also more complicated.

Alternatively, one might approximate the demand for each demand class by an independent Brownian motion process, i.e., $D_i(s, t)$ is normally distributed with mean and standard deviation given by $\mu_i(t - s)$

and $\sigma_i\sqrt{t-s}$. As this process has independent increments, we can apply model (2)–(5), but some care is needed in the subsequent analysis. As with the case of compound Poisson demand, we would have to adapt the conditional distribution of $B_{i,i-1}$ in (11), as it is no longer binomial. One would also need to examine how best to specify and measure service, when demand is approximated by a Brownian motion process.

8.4. Future Extensions

There are a number of questions and issues left for future research. One is to understand better how our allocation process performs for various problem criteria. We would also like to extend our analyses for alternative allocation mechanisms. Second, we assume that inventory shortages result in backorders; we have not found an easy way to modify the current model to accommodate a lost-sales assumption, but view this as a research opportunity. Finally, we develop our analysis and solution procedure for one specification of the inventory problem, namely, the SLP. One might consider another problem specification whereby we minimize the total inventory-related costs, with no service-level constraints. Questions include: What is a good algorithm for determining the minimum-cost policy? Are there any structural results that can be exploited for solving this problem?

9. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at <http://mansci.journal.informs.org/>.

Acknowledgments

The authors thank the associate editor and the two referees for their very helpful feedback on earlier versions. This research has been supported in part by the MIT Leaders for Manufacturing Program, a partnership between MIT and global manufacturing firms; and by the Singapore-MIT Alliance, an engineering education and research collaboration among the National University of Singapore, Nanyang Technological University, and MIT.

Appendix

We argue here the sample-path equivalence between the FCFS rule and the threshold clearing rule (Deshpande et al. 2003, Deshpande and Cohen 2005). We will do this for a two-class system, but the result extends to N classes.

To see the equivalence between the FCFS rule and the threshold clearing rule, we need to examine how each rule allocates a replenishment quantity Q . We describe this first for the FCFS rule and then discuss the correspondence to the threshold clearing rule.

The shortfall for class 1 is defined by Equation (1) as

$$SF_1(t) = [c_1 - IOH]^+ + B_{1,1}(t).$$

We can show that the shortfall can always be expressed as

$$SF_1(t) = D_1(u, t), \quad (18)$$

for some value of u , $u \leq t$, where $D_i(s, t)$ denotes the demand for class i over the interval $(s, t]$. Furthermore, we can show that for any time t that the class-2 backorders are given by

$$B_{2,2}(t) = D_2(u, t) \quad (19)$$

with the same value of u , $u \leq t$.

Suppose that a replenishment quantity Q arrives at time t . We have two cases to consider.

Case 1. If $Q \geq B_{2,2}(t) + SF_1(t)$: The replenishment quantity is sufficient to eliminate all backorders and restore the on-hand inventory above the critical level for class 1. Thus, denoting t^+ as the time epoch after the allocation, we have

$$B_{2,2}(t^+) = D_2(t, t) = 0,$$

$$SF_1(t^+) = D_1(t, t) = 0,$$

$$IOH(t^+) = Q - B_{2,2}(t^+) - SF_1(t^+) + c_1 \geq c_1.$$

Case 2. If $Q < B_{2,2}(t) + SF_1(t)$: The replenishment quantity is not sufficient to eliminate the class-2 backorders and the class-1 shortfall. By definition, the FCFS rule combines the class-2 backorders with the class-1 shortfalls and will fill these in the order of occurrence. As a consequence, this allocation rule will fill all class-2 backorders and class-1 shortfalls that were incurred prior to some time epoch; if we let v denote this time epoch, then after the allocation we still have class-2 backorders and a class-1 shortfall, where each corresponds to the demand subsequent to the time epoch v . That is,

$$B_{2,2}(t^+) = B_{2,2}(t) - Q_2 = D_2(v, t),$$

$$SF_1(t^+) = SF_1(t) - Q_1 = D_1(v, t),$$

where $Q_1 + Q_2 = Q$. By substituting (18) and (19) into the above equations, we see that the replenishment quantity Q is divided as follows:

$$\begin{aligned} Q_2 &= D_2(u, v), \\ Q_1 &= D_1(u, v). \end{aligned} \quad (20)$$

The inventory on-hand after the replenishment is given by

$$IOH(t^+) = (c_1 - SF_1(t^+))^+,$$

which reflects the fact that we allocate Q_1 to fill all class-1 backorders before rebuilding the class-1 reserve stock.

We now consider the threshold clearing rule from Deshpande et al. (2003).

For the first case, when $Q \geq B_{2,2}(t) + SF_1(t)$, the threshold clearing rule will do exactly the same as the FCFS rule; it will eliminate all of the backorders for both classes.

For the second case, when $Q < B_{2,2}(t) + SF_1(t)$, the threshold clearing rule identifies two time epochs, denoted as t_{B_2} and t_{K_j} in Deshpande et al. (2003) with $t_{B_2} < t_{K_j} < t$. The quantity Q is then allocated similar to (20): $Q_2 = D_2(t_{B_2}, t_{K_j})$, $Q_1 = D_1(t_{B_2}, t_{K_j})$. This allocation clears all back-ordered demand that arrived prior to t_{K_j} ; the remaining inventory is first used to clear any remaining class-1 backorders and then held as on-hand inventory. The first time epoch t_{B_2} is the most recent epoch at which the on-hand inventory reaches the class-1 critical level and corresponds

to time epoch u in (20); in the parlance of our SSS, it is the time at which the stage-2 on-hand inventory reaches zero, after which it incurs both external and internal backorders. The second time t_{K_j} is defined as the demand epoch at which there have been Q demands since time t_{B_2} ; that is, for given t_{B_2} , we specify t_{K_j} such that $D(t_{B_2}, t_{K_j}) = \bar{D}_1(t_{B_2}, t_{K_j}) + D_2(t_{B_2}, t_{K_j}) = Q$. This second time epoch corresponds to time epoch v in (20).

References

- Atkins, D., K. Katircioglu. 1996. Managing inventory for multiple customers requiring different levels of service. Working paper, University of British Columbia, Vancouver, BC, Canada.
- Dekker, R., M. J. Kleijn, P. J. de Rooij. 1998. A spare parts stocking policy based on equipment criticality. *Int. J. Production Econom.* **56–57** 69–77.
- Dekker, R., R. M. Hill, M. J. Kleijn, R. H. Teunter. 2002. On the (S-1, S) lost sales inventory model with priority demand classes. *Naval Res. Logist.* **49** 593–610.
- Deshpande, V., M. A. Cohen. 2005. A nested threshold inventory rationing policy for multiple demand classes in inventory systems with replenishment. Working paper, Krannert School of Management, Purdue University, West Lafayette, IN.
- Deshpande, V., M. A. Cohen, K. Donohue. 2003. A threshold inventory rationing policy for service-differentiated demand classes. *Management Sci.* **49**(6) 683–703.
- de Véricourt, F., F. Karaesmen, Y. Dallery. 2000. Dynamic scheduling in a make-to-stock system: A partial characterization of optimal policies. *Oper. Res.* **48**(5) 811–819.
- de Véricourt, F., F. Karaesmen, Y. Dallery. 2002. Optimal stock allocation for a capacitated supply system. *Management Sci.* **48**(11) 1486–1501.
- Evans, R. V. 1968. Sales and restocking policies in a single inventory system. *Management Sci.* **14**(7) 463–473.
- Fadiloglu, M. M., Ö. Bulut. 2005. An embedded Markov chain approach to the analysis of inventory systems with backordering under rationing. Working paper, Department of Industrial Engineering, Bilkent University, Ankara, Turkey.
- Frank, K. C., R. Q. Zhang, I. Duenyas. 2003. Optimal policies for inventory systems with priority demand classes. *Oper. Res.* **51**(6) 993–1002.
- Graves, S. C. 1985. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Sci.* **31**(10) 1247–1256.
- Graves, S. C. 1996. A multi-echelon inventory model with fixed replenishment intervals. *Management Sci.* **42**(1) 1–18.
- Ha, A. Y. 1997a. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Sci.* **43**(8) 1093–1103.
- Ha, A. Y. 1997b. Stock rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Res. Logist.* **44**(5) 457–472.
- Kaplan, A. 1969. Stock rationing. *Management Sci.* **15**(5) 260–267.
- Kleijn, M. J., R. Dekker. 1998. An overview of inventory systems with several demand classes. Econometric Institute Report 9838/A, Erasmus University, Rotterdam, The Netherlands.
- Melchior, P. 2001. Rationing policies for an inventory model with several demand classes and stochastic lead times. Working paper, Department of Operations Research, University of Aarhus, Denmark.
- Melchior, P., R. Dekker, M. J. Kleijn. 2000. Inventory rationing in an (s, Q) inventory model with lost sales and two demand classes. *J. Oper. Res. Soc.* **51**(1) 111–122.
- Moon, I., S. Kang. 1998. Rationing policies for some inventory systems. *J. Oper. Res. Soc.* **49**(5) 509–518.
- Nahmias, S., S. Demmy. 1981. Operating characteristics of an inventory system with rationing. *Management Sci.* **27**(11) 1236–1245.
- Topkis, D. 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Sci.* **15**(3) 160–176.
- Veinott, A. F. 1965. Optimal policy in a dynamic, single product, nonstationary inventory model with several demand classes. *Oper. Res.* **13**(5) 761–778.
- Zipkin, P. H. 2000. *Foundations of Inventory Management*. McGraw-Hill, New York.