MIT Sloan School of Management

MIT Sloan Working Paper 4542-05
May 2005

A Single-Product Inventory Model for Multiple Demand Classes

Hasan Arslan, Stephen C. Graves, and Thomas Roemer

A Single-Product Inventory Model for Multiple Demand Classes[1]

Hasan Arslan,[2] Stephen C. Graves,[3] and Thomas Roemer[4]

March 5, 2005

Abstract

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 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 to embed within a multi-echelon setting.

[1] 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.

[2] Sawyer School of Management, Suffolk University, Boston MA, 02108-2770, harslan@suffolk.edu

[3] Leaders for Manufacturing Program and A. P. Sloan School of Management, MIT, Cambridge MA 02139-4307, sgraves@mit.edu

[4] Leaders for Manufacturing Program and A. P. Sloan School of Management, MIT, Cambridge MA 02139-4307, troemer@mit.edu

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 twenty-four hours, while a "bronze contract" promises a 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.

The key contribution of this paper is to show how to map this problem into a serial inventory system. This mapping facilitates the characterization of the steady-state behavior of the inventory system. We then develop an approximate solution procedure to the so-called Service Level Problem; that is, we want to find the critical-level policy that meets specified fill-rate targets for each demand class with the least inventory. We show with both bounds and a numerical experiment that this heuristic is quite robust and near optimal. We also 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.

We have organized the paper into seven 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 section three, we show how to map this inventory system into a serial inventory system. In section four, based on this mapping, 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. In section five, we pose the Service Level Problem, in which we minimize the expected inventory while satisfying a service level requirement for each demand class. Furthermore, we provide a heuristic solution approach for the Service Level Problem. In section six, we provide a numerical experiment both to compare our proposed heuristic with the optimal solution and to show the value from rationing. In the final section we discuss possible extensions and directions for future research.

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 divide the research based on the assumed control policy, periodic or continuous review, and on assumed 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.

	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) Deshpande et al. (2003) Dekker et al. (1998)
N demand classes	Veinott (1965) Topkis (1968)	Katircioglu and Atkins (1996)	Melchiors (2001) Dekker et al. (2000)	

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

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, Katircioglu and Atkins (1996) and Frank et al. (2003) analyze periodic-review inventory systems with multiple stochastic demand classes. Katircioglu and Atkins (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 an (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 without restrictions on the number of orders outstanding. 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.

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 inventory level drops to the critical level. Melchiors (2001) extend the model in Melchiors et al. (2000) to multiple Poisson demand classes with stochastic replenishment lead-times. Moreover, he considers a non-stationary 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 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. (2000) extends the model in Dekker et al. (1998) to multiple demand classes with stochastic replenishment lead-times, but switching to a lost sales environment rather than allowing backorders. They assume one-for-one replenishment policy and a critical-level policy to ration inventory among demand classes. In a lost sales environment, handling incoming replenishment orders is not a dilemma; each incoming replenishment order simply replenishes the inventory. They develop numerical solution methods to efficiently calculate the optimal base stock level and critical levels with or without service level constraints.

Ha (1997a) considers a make-to-order production system with a single production facility and multiple demand classes for the end product. He assumes a lost sales environment, exponentially distributed production time, and Poisson demand for each demand class. He shows that a stationary critical level policy is optimal. Ha (1997b) extends the study in Ha (1997a) by allowing backorders to occur. Vericourt et al. (2000, 2002) consider the multiple-demand class extension of the two-demand class study in Ha (1997a). They develop a characterization of the optimal policy for the backorders case with zero set-up 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;

(iv) We backorder any demand that is not immediately satisfied from on-hand inventory.

In addition to these assumptions, we need to describe how we will 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 $\boldsymbol{c} = \left\{ c_1, c_2, \cdots, c_{N-1} \middle| c_i \in Z^+ \cup \{0\} \text{ and } c_{i-1} \le c_i \right\}$. We stop serving demand class-*i* once the on-hand inventory reaches or falls below the critical stock level c_{i-1}; by assumption, we then backorder all demand for class-*i* until the on-hand inventory is raised above c_{i-1}. For class-*1*, we set its critical level $c_0 = 0$; thus, we continue to fill class-*1* demand until the on-hand inventory is completely depleted, at which point we backorder any subsequent class-*1* demand.

We define $s_i = c_i - c_{i-1}$ for i=1,…N-1 to be the *reserve stock* for class-*i*, as it represents the quantity of stock that we protect or reserve for this demand class. By the definition of the critical-level policy, each of these reserve stocks is non-negative. For stage N, we define the reserve stock $s_N = R - c_{N-1}$, for which we require no assumptions about its sign.

We also need an assumption with regard to how we allocate an inventory replenishment at the time it is received. The primary issue is to decide how much of the replenishment we use to fill backorders versus use to re-build the reserve stock for higher-priority demand classes. We defer until the next section the presentation and discussion of our allocation assumption, as it will be easier to explain in the context of the problem mapping to a serial inventory system.

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. To ease the presentation, we denote the former as the DCS (demand-class system) and the latter as the SSS (serial stage system).

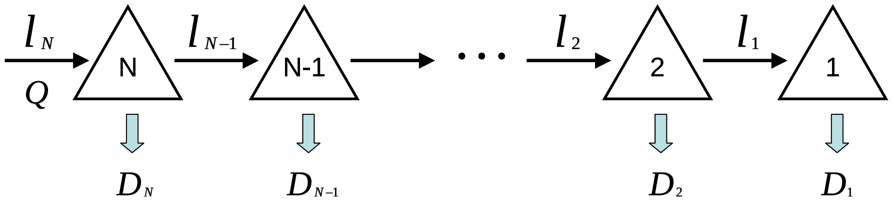

Figure 1: Serial Inventory System with Demand at Each Installation

We consider an N-stage serial system (SSS), as shown in Figure 1, and assume that it operates as follows:

(i) Each stage *i*, $i \in \{|1, N-1|\}$ operates with a one-for-one continuous review base-stock policy with non-negative base-stock level s_i., and is replenished by its upstream stage *i+1* with replenishment lead time *l=0*.

(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 $s_N = R - c_{N-1}$.

(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.

We contend that this SSS is equivalent to the DCS, as described in the prior section. To establish this equivalence, we will show 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 $c_{j-1} < IOH \le c_j$ 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*, $i \in \left\{\left|1, j-1\right|\right\}$ equals its base stock *s*$_i$. For each stage *i*, $i \in \left\{\left|j+1, N\right|\right\}$, there is no on-hand inventory, while stage j has on-hand inventory equal to $IOH - c_{j-1}$. If there were a demand for stage $i \in \left\{\left|1, j\right|\right\}$, the serial system fills the demand, the on-hand inventory at each stage *i*, $i \in \left\{\left|1, j-1\right|\right\}$ remains at its base stock *s*$_i$, and the on-hand inventory at stage *j* (as well as the *IOH*) is depleted by one. However, if there were a demand for stage *i*, $i \in \left\{\left|j+1, N\right|\right\}$, 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 $i \in \left\{\left|1, N-1\right|\right\}$ 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 a replenishment arrives. We will do this by first describing our assumptions for the SSS and then interpreting how these assumptions apply to the DCS.

Consider the SSS with on-hand inventory *IOH* in the system at time *t*. Suppose that $c_{j-1} < IOH \leq c_j$ for some $j \in \left\{\left|1, N\right|\right\}$. The on-hand inventory equals its base stock *s*$_i$ for stage $i \in \left\{\left|1, j-1\right|\right\}$, is equal to $IOH - c_{j-1}$ for stage *j*, and is zero for stage $i \in \left\{\left|j+1, N\right|\right\}$. There are no backorders at each stage *i*, $i \in \left\{\left|1, j\right|\right\}$. Each stage *i*, $i \in \left\{\left|j+1, N\right|\right\}$, has backorders given by:

$$B_{ii}\left(t\right) = D_i\left(\tau_i, t\right) \text{ and } B_{i,i-1}\left(t\right) = \sum_{k=1}^{i-1} D_k\left(\tau_i, t\right)$$

where $D_i\left(s,t\right)$ is the external demand during time interval (*s, t*] at stage *i*; $B_{ii}\left(t\right)$ is the number of backorders at time *t* at stage *i*, due its external demand; $B_{i,i-1}\left(t\right)$ is the number of backorders at time *t* at stage *i*, due to internal demand from stage *i-1*; and τ_i is the most recent time at which stage *i* stocks out. Stage *i* stocks out once the on-hand inventory reaches c_{i-1}; thus, we determine τ_i to be the most recent time epoch at which $IOH = c_{i-1}$.

Suppose stage *N* of the SSS receives a replenishment of Q at time *t*. There are two cases to consider.

$\underline{Q \ge B_{NN}\left(t\right) + B_{N,N-1}\left(t\right)}$: The replenishment quantity is sufficient to fill all backorders at stage *N*, as well as all downstream stages. Thus, this replenishment returns the on-hand inventory at each stage *i*, $i \in \left\{\left|1, N-1\right|\right\}$ to its base stock s_i; any remaining on-hand inventory is held at stage *N*.

$\underline{Q < B_{NN}\left(t\right) + B_{N,N-1}\left(t\right)}$: The replenishment quantity is not sufficient to fill all backorders at stage *N*, and we need to decide how to allocate the inventory between the two types of backorders. We assume that we fill these backorders in the order of occurrence, with no differentiation between external and internal backorders.

In particular, we find the earliest time $s : \tau_N < s < t$ such that $\sum_{k=1}^{N} D_k\left(\tau_N, s\right) = Q$. We set the replenishment quantities to be $Q_{NN} = D_N\left(\tau_N, s\right)$ and $Q_{N,N-1} = \sum_{k=1}^{N-1} D_k\left(\tau_N, s\right)$. Thus, after this allocation the remaining backorders at stage N are $B_{NN}\left(t^+\right) = D_N\left(s,t\right)$ and $B_{N,N-1}\left(t^+\right) = \sum_{k=1}^{N-1} D_k\left(s,t\right)$.

We assume this process repeats at each downstream stage. For instance, stage *N-1* receives the replenishment of $Q_{N,N-1}$. If it is sufficient to cover the backorders at stage *N-1*, then we fill these backorders and hold the remainder on hand. If the replenishment is not sufficient to cover the backorders, then we allocate $Q_{N,N-1}$ to fill these backorders in their order of occurrence, as described above. We repeat this allocation process at each downstream stage, until we reach a stage at which the stage's replenishment covers the backorders at the stage or we reach stage 1.

We assume the same allocation process for the DCS. Namely, we assume that we fill backorders in the order of occurrence, with no differentiation between external and internal backorders.

We start with demand class *N*, and need to decide how to split the replenishment quantity between class-*N* backorders and the outstanding replenishment requests from stages *i*, $i \in \{|1, N-1|\}$. These outstanding replenishment requests can entail both customer backorders and replenishments to re-build the reserve stock at the higher-priority classes. If the replenishment quantity is not sufficient, we will allocate it in the order of the demand occurrences that created the backorders or replenishment requests. This process repeats with each demand class until we reach a class for which the replenishment covers the backorders at the class or we reach demand class 1.

This allocation scheme is not optimal. However, it seems to be reasonable given that it allocates the inventory at each stage (or demand class) to the internal and external backorders at the stage in the order of occurrence. Thus, at each stage it tries to balance servicing external versus internal backorders, where the internal backorders include the re-building of a reserve stock at downstream stages (higher priority demand classes). This 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. As in Graves (1996), this scheme permits significant tractability in the analysis of the inventory system, as will be seen.

This completes the discussion equating the DCS to an SSS. We find this equivalence to be helpful in visualizing the operation of the N-demand-class inventory system, 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 backorders at time *t* at stage *i*;

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 $[\,\cdot\,]^+$ denotes the positive part of the expression, l_i is the replenishment lead-time for stage *i*, and $i \in \{|1, N|\}$:

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

$$B_i(t) = \left[-IL_i(t)\right]^+ \tag{2}$$

$$B_i(t) = B_{ii}(t) + B_{i,i-1}(t) \tag{3}$$

$$B_{1,0}(t) = 0; B_{N+1,N}(t) = 0 \tag{4}$$

The explanation for (1) 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. All items that were in-process at time *t* will arrive at stage *i* by time $t + l_i$, by the definition of the lead-time. However, none of the backorders at stage *i* + 1 at time *t* can arrive to stage *i* by time $t + l_i$, again by the definition of the lead-time. 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 (2) we state the backorders to be the negative part of the inventory level. In equation (3) 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 (4), 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 $\left[s_N + 1, s_N + Q\right]$, given the assumption of a reorder-point, reorder-quantity replenishment system with these parameters (Zipkin 2000, p. 193).

We now use these observations to re-write the steady-state form for equations (1) – (4):

$$IL_N = IP_N - \sum_{i=1}^{N} D_i^L \tag{5}$$

$$IL_i = s_i - B_{i+1,i} \quad for\ i = 1, 2, \cdots N\text{ -}1 \tag{6}$$

$$B_i = \left[-IL_i\right]^+ \tag{7}$$

$$B_i = B_{ii} + B_{i,i-1} \tag{8}$$

$$B_{1,0} = 0 \tag{9}$$

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 (5) – (9) to determine the steady-state distribution of the inventory level at each stage. We intend to use (5) or (6) to find the distribution of the inventory level, and then (7) 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 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 $j \in \left\{\left|0, n\right|\right\}$ that

$$\Pr\left[B_{i,i-1} = j \,\middle|\, B_i = n\right] = \binom{n}{j} p_i^j \left(1 - p_i\right)^{n-j} \qquad \text{where } p_i = \frac{\sum_{j=1}^{i-1} \lambda_j}{\sum_{j=1}^{i} \lambda_j}\ . \tag{10}$$

As 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 $\left(n, p_i\right)$. 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 $B_i\left(t\right) = \sum_{j=1}^{i} D_j\left(s,t\right)$ for some value of s<t. Due to the memory-less 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 $\left(Q,R\right)$ and $\left(c_1,\ldots,c_{N-1}\right)$. The procedure starts from the most upstream stage *N* and moves iteratively to each downstream stage, as follows:

Step 1: *Set i = N. Determine the steady-state distribution of IL_N.* We obtain the distribution of IL_N from equation (5) by convolving the distribution of IP_N with that for $\sum_{i=1}^{N} D_i^L$. The former is a uniform random variable on the interval $\left[s_N + 1, s_N + Q\right]$; the latter is a Poisson random variable with mean $L\sum_{i=1}^{N}\lambda_i$.

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

Step 3: *Determine the steady-state distribution of $B_{i,i-1}$.* We use the distribution for B_i with (10) to get the un-conditioned distribution for $B_{i,i-1}$.

Step 4: *Set $i := i-1$. Determine the steady-state distribution of IL_i from (6).*

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 critical levels $\left\{c_i : i = 1,\ldots, N-1\right\}$ (or equivalently the reserve stocks $\left\{s_i = c_i - c_{i-1} : i = 1,\ldots, N-1\right\}$). 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 tradeoff 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 promises or 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:

$$\textbf{\textit{SLP}} \quad Min\ z = \sum_{i=1}^{N} E\left[IL_i\right]^+$$

$$\begin{aligned} s.t. \quad & Fillrate_i \geq \beta_i && for\ \ i = 1, \ldots, N \\ & s_i \geq 0,\ integer && for\ \ i = 1, \ldots, N-1 \\ & s_N\ integer && \end{aligned}$$

where

$$Fillrate_i = \begin{Bmatrix} \Pr\left(IL_i > 0\right) & if\ \ s_i > 0 \\ Fillrate_{i+1} & if\ \ s_i = 0 \end{Bmatrix}\ for\ \ i = 1, \ldots, N-1$$

$$Fillrate_N = \Pr\left(IL_N > 0\right)$$

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 pre-specified.

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 demand class. If the reserve stock for the demand class is positive, then for Poisson demand the fill rate equals the probability that the inventory level is positive. When the reserve stock for the demand class *i* is zero, then the fill rate for demand class *i* is the same as that for demand class *i+1*. This is because when s_i = 0, there is no distinction in order fulfillment between a demand from class *i* and a demand from class *i+1*.

In formulating the SLP, we expect (although don't require) that the higher-priority demand classes have larger fill-rate targets; that is, we expect $\beta_1 \geq \beta_2 \geq \ldots \geq \beta_N$. 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 the model (5) – (9), 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 $\left(s_i,\ldots,s_N\right)$. In the following, we will at times use the notation $IL_i\left(s_i,\ldots,s_N\right)$ to make this dependence explicit.

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.

SPA uses the model given by (5) - (9) to find the reserve stock for each demand class sequentially, starting with stage *N*. For each demand class *i*, SPA finds the minimum value for its reserve stock that satisfies its fill-rate target, given the previously-determined reserve stocks for demand classes *i+1,…,N*. We state the algorithm as follows:

1. Find reserve stock and fill rate for stage *N*: $\hat{s}_N = \min\left\{s : \Pr\left(IL_N\left(s\right) > 0\right) \geq \beta_N\right\}$; and $Fillrate_N = \Pr\left(IL_N\left(\hat{s}_N\right) > 0\right)$; let *i := N-1*.
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 $Fillrate_{i+1} < \beta_i$: $\hat{s}_i = \min\left\{s : \Pr\left(IL_i\left(s,\hat{s}_{i+1},\ldots,\hat{s}_N\right) > 0\right) \geq \beta_i\right\}$; and $Fillrate_i = \Pr\left(IL_i\left(\hat{s}_i,\hat{s}_{i+1},\ldots,\hat{s}_N\right) > 0\right)$
3. Stop if *i:= 1*. Otherwise, let *i := i-1* and repeat step 2.

The Single-Pass-Algorithm yields a feasible solution for the SLP by construction: at each iterative step, 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 for the SPA should be quite good. To develop this argument, it will be helpful to re-write the objective function of SLP as

$$z = \sum_{i=1}^{N} E\left[IL_i\right]^+ = \sum_{i=1}^{N}\left(s_i + E\left[B_{ii}\right]\right) + \frac{Q+1}{2} - L\sum_{i=1}^{N}\lambda_i. \tag{11}$$

We obtain this expression from substituting (5), (6) and (8)

into $E\left[IL_i\right]^{+} = E\left[IL_i\right] + E\left[B_i\right]$, with the observation that $E\left[IP_N\right] = s_N + \frac{Q+1}{2}$. Thus, we observe that the objective function consists of the sum of the reserve stocks and the sum of the external backorders, plus a constant *K*:

$$z = \sum_{i=1}^{N} s_i + \sum_{i=1}^{N} E\left[B_{ii}\right] + K \ . \tag{12}$$

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.

We first show that moving one unit of reserve stock from class j to class j-1 cannot decrease the inventory level at any of the higher-ranked classes, but can result in more backorders.

Proposition 1: Consider two stocking policies $\left(s_1^1, s_2^1, \ldots, s_N^1\right)$ and $\left(s_1^2, s_2^2, \ldots, s_N^2\right)$ 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) $IL_i\left(s_i^2, \ldots, s_N^2\right) \geq IL_i\left(s_i^1, \ldots, s_N^1\right)$ for $i = 1, 2, \ldots j-1$, and

(ii) $\sum_{i=1}^{j} B_{ii}\left(s_i^2, \ldots, s_N^2\right) \geq \sum_{i=1}^{j} B_{ii}\left(s_i^1, \ldots, s_N^1\right)$.

Proof.

From (5) and (6) we find that $IL_j\left(s_j^2, \ldots, s_N^2\right) = IL_j\left(s_j^1 - 1, s_{j+1}^1, \ldots, s_N^1\right) = IL_j\left(s_j^1, \ldots, s_N^1\right)$-1.

Thus from (7) and (8), we obtain $B_j\left(s_j^2, \ldots, s_N^2\right) \leq B_j\left(s_j^1, \ldots, s_N^1\right) + 1$

and $B_{j,j-1}\left(s_j^2, \ldots, s_N^2\right) \leq B_{j,j-1}\left(s_j^1, \ldots, s_N^1\right) + 1$. We can now use this result in (6) to show that:

$$\begin{aligned} IL_{j-1}\left(s_{j-1}^2, \ldots, s_N^2\right) &= s_{j-1}^2 - B_{j,j-1}\left(s_j^2, \ldots, s_N^2\right) \\ &\geq s_{j-1}^1 + 1 - B_{j,j-1}\left(s_j^1, \ldots, s_N^1\right) - 1 = IL_{j-1}\left(s_{j-1}^1, \ldots, s_N^1\right) \end{aligned}$$

From $IL_{j-1}\left(s_{j-1}^2, \ldots, s_N^2\right) \geq IL_{j-1}\left(s_{j-1}^1, \ldots, s_N^1\right)$ and $s_i^2 = s_i^1, i = 1, \ldots j-2$, we find that $IL_i\left(s_i^2, \ldots, s_N^2\right) \geq IL_i\left(s_i^1, \ldots, s_N^1\right)$ for $i = 1, 2, \ldots j-2$, which proves the first result.

For the second result, we can make a sample path comparison. Suppose at time t, we have $B_{j+1,j}\left(t\right) < s_j^1$; then there are no backorders for either case:

$\sum_{i=1}^{j} B_{ii}\left(t \middle| s_i^2, \ldots, s_N^2\right) = \sum_{i=1}^{j} B_{ii}\left(t \middle| s_i^1, \ldots, s_N^1\right) = 0$. If $B_{j+1,j}\left(t\right) \geq s_j^1$, then

$B_j\left(t\left|s_j^2,\ldots,s_N^2\right.\right)=B_j\left(t\left|s_j^1,\ldots,s_N^1\right.\right)+1$ and either $B_{jj}\left(t\left|s_j^2,\ldots,s_N^2\right.\right)=B_{jj}\left(t\left|s_j^1,\ldots,s_N^1\right.\right)+1$ or $B_{j,j-1}\left(t\left|s_j^2,\ldots,s_N^2\right.\right)=B_{j,j-1}\left(t\left|s_j^1,\ldots,s_N^1\right.\right)+1$. In the former case we have $\sum_{i=1}^{j}B_{ii}\left(t\left|s_i^2,\ldots,s_N^2\right.\right)=\sum_{i=1}^{j}B_{ii}\left(t\left|s_i^1,\ldots,s_N^1\right.\right)+1$; in the latter case we have $\sum_{i=1}^{j}B_{ii}\left(t\left|s_i^2,\ldots,s_N^2\right.\right)=\sum_{i=1}^{j}B_{ii}\left(t\left|s_i^1,\ldots,s_N^1\right.\right)$. This proves the result. □

We now use these results to establish bounds on the optimal solution. We first show that the solution for the SPA provides a bound on the sum of the reserve stock.

Proposition 2: For all feasible solutions $\left(s_1,\ldots,s_N\right)$ for the SLP, we have $\sum_{i=j}^{N}s_i\geq\sum_{i=j}^{N}\hat{s}_i$ for all *j* where $\left(\hat{s}_1,\ldots,\hat{s}_N\right)$ is the solution found by the SPA.

Proof.

Suppose we have a feasible solution $\left(s_1^1,s_2^1,\ldots,s_N^1\right)$ such that $\sum_{i=j}^{N}s_i^1<\sum_{i=j}^{N}\hat{s}_i$. We will iteratively construct a series of feasible solutions, which leads to a contradiction of the supposition.

In order for $\left(s_1^1,s_2^1,\ldots,s_N^1\right)$ to be a feasible solution, we must have $s_N^1\geq\hat{s}_N$; otherwise the fill-rate constraint for class *N* is violated. If $s_N^1=\hat{s}_N$ then we must have $s_{N-1}^1\geq\hat{s}_{N-1}$ by the same logic. If both $s_N^1=\hat{s}_N$, $s_{N-1}^1=\hat{s}_{N-1}$, then we must have $s_{N-2}^1\geq\hat{s}_{N-2}$ and so on.

Iterative Step: Let *k* be the largest index such that $s_k^1>\hat{s}_k$; that is, $s_N^1=\hat{s}_N,\ldots,s_{k+1}^1=\hat{s}_{k+1}$, and $s_k^1>\hat{s}_k$. If *k ≤ j*, we have a contradiction of the original supposition that $\sum_{i=j}^{N}s_i^1<\sum_{i=j}^{N}\hat{s}_i$. If there does not exist an index *k*, then we must have $\left(s_1^1,s_2^1,\ldots,s_N^1\right)=\left(\hat{s}_1,\hat{s}_2,\ldots,\hat{s}_N\right)$, which is also a contradiction of the original supposition. Given *k > j*, then we construct a new solution: $s_k^2=s_k^1-1, s_{k-1}^2=s_{k-1}^1+1,$ and $s_i^2=s_i^1,\forall i\neq k,k-1$. By application of proposition 1 (i), we can show that this new solution is feasible. Since *k > j*, we have that $\sum_{i=j}^{N}s_i^1=\sum_{i=j}^{N}s_i^2<\sum_{i=j}^{N}\hat{s}_i$.

We now use the new solution to repeat the *Iterative Step*. At each step we move one unit of reserve stock from a higher-numbered class to a lower-numbered class to create a new feasible solution. The number of possible iterative steps is finite as each unit of reserve stock can be moved at most *N-1* times. Therefore, at some step *n* we have either $s_i^n = \hat{s}_i$ for $i = j, \dots, N$ or $s_j^n > \hat{s}_j$ and $\left(s_{j+1}^n, \dots, s_N^n\right) = \left(\hat{s}_{j+1}, \dots, \hat{s}_N\right)$, both of which are contradictions of the original supposition. □

As a special case of this proposition, we see that $\sum_{i=1}^{N} \hat{s}_i$ is a lower bound on the sum of the reserve stocks in the objective function of the SLP. We denote $\sum_{i=1}^{N} \hat{s}_i = \hat{R}$, since the sum of the reserve stocks is the reorder point for the critical-level policy.

Proposition 3: Consider two stocking policies $\left(s_1^1, s_2^1, \dots, s_N^1\right)$ and $\left(s_1^2, s_2^2, \dots, s_N^2\right)$ with $\sum_{i=1}^{j} s_i^1 \le \sum_{i=1}^{j} s_i^2$ for $j = 1, \dots, N-1$ and $\sum_{i=1}^{N} s_i^1 = \sum_{i=1}^{N} s_i^2$. Then we have $z\left(s_1^1, \dots, s_N^1\right) \le z\left(s_1^2, \dots, s_N^2\right)$.

Proof.

From (12) and the assumption that $\sum_{i=1}^{N} s_i^1 = \sum_{i=1}^{N} s_i^2$, we need to show that $\sum_{i=1}^{N} E\left[B_{ii}\left(s_i^1, \dots s_N^1\right)\right] \le \sum_{i=1}^{N} E\left[B_{ii}\left(s_i^2, \dots s_N^2\right)\right]$. This result follows directly from application of proposition 1 (ii). Starting with the stocking policy $\left(s_1^2, s_2^2, \dots, s_N^2\right)$, we can construct a series of new policies in which we move one unit of reserve stock from class *j-1* to class *j*, and eventually reach the stocking policy $\left(s_1^1, s_2^1, \dots, s_N^1\right)$. From proposition 1 (ii), each such move reduces the external backorders in classes *1 ,2, … j*, and has no impact on backorders at class *j+1,…, N*. Thus we show that $\sum_{i=1}^{N} B_{ii}\left(s_i^1, \dots s_N^1\right) \le \sum_{i=1}^{N} B_{ii}\left(s_i^2, \dots s_N^2\right)$, and we get the desired result by taking expectations. □

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

$$\hat{R}+\sum_{i=1}^{N} E\left[B_{ii}\left(s_i=0, s_{i+1}=0,\ldots s_{N-1}=0, s_N=\hat{R}\right)\right]+K \tag{13}$$

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 the next section 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 $L=0.25$ years (3 months), and the reorder quantity is $Q=1$. The fill-rate targets are: β_1 = 0.99, β_2 = 0.94, and β_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, we stop serving demand classes *3* and *2* once the on-hand inventory drops to 3 and to 2, respectively. We can use (11) to break the objective value into its constituent parts:

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

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 = 7.08. The only difference between the optimal solution and the SPA solution is that one unit of reserve stock has been moved from demand class *1* and demand class *2* to demand class *3*; this move reduces the backorders at demand class *3* without jeopardizing the fill-rate constraint for demand class *1* and for demand class *2*. The move does 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 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 27% higher than the optimal solution.

Suppose now that the fill-rate targets are $\beta_1 = 0.99$, $\beta_2 = 0.93$, and $\beta_3 = 0.70$. The SPA finds the solution: $\hat{s}_1 = 2,\ \hat{s}_2 = 2,\ \hat{s}_3 = 10$ with z = 6.24. The expected backorders for the SPA solution is 0.24, and the lower bound on the expected backorders is 6.04. The optimal solution entails moving a unit of reserve stock from demand class 1 to demand class 3: $s_1 = 1,\ s_2 = 2,\ s_3 = 11$ with z = 6.14.

In each of these two cases we see that the SPA generates a near-optimal solution. We are able to improve slightly the SPA solution by shifting a unit of reserve stock from a higher-ranked class to a lower-ranked class. The result of this shift is that we reduce some backorders at the lower-ranked demand class, yet still satisfy the fill-rate target at the higher-ranked class.

6. Numerical Experiment:

To test the effectiveness of the SPA on the SLP, we compare its solution to the optimal solution on two set of test problems in two experiments. For the first experiment, we examine the performance of 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, or 18. The replenishment lead-time from the outside supplier is one of the three values: L = 1/24 year, ¼ year, ½ year. There are three possible values for the fill-rate target for each of the demand classes: $\beta_1 = 0.9,\ 0.95, \text{ or } 0.99,\ \ \beta_2 = 0.8,\ 0.9, \text{ or } 0.95,\ \text{ and }\ \beta_3 = 0.7,\ 0.8, \text{ or } 0.9$. 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 (1 candidate), the replenishment lead-time (3 candidates), the reorder quantity (4 candidates), the set of desired fill-rates (20 candidates), and the set of demand rates (4 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 (12); the lower bound from equation (13); 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 total reserve stock equal to $\hat{s}_1 + \hat{s}_2 + \hat{s}_3 + 1$. 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 guarantee that the total reserve stock 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: $s_3 \geq \hat{s}_3; s_2 + s_3 \geq \hat{s}_2 + \hat{s}_3;$ and $s_1 + s_2 + s_3 = \hat{s}_1 + \hat{s}_2 + \hat{s}_3$.

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 SPA is 3.24%.

In Table 2, 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 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 sub-optimal 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 this set of 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 = ¼ year and the reorder quantity Q = 4 for each test problem. In Table 3, we specify the fill-rate targets and the demand rates for each test problem. As with the first experiment, 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 3. 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.

Lead Time	**L = 1/24** **0.52%**	**L = ¼** **0.66%**	**L= ½** **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	$0.05 \le \beta_1 - \beta_3 < 0.15$ **0.32%**	$0.15 \le \beta_1 - \beta_3 < 0.25$ **0.56%**	$0.25 \le \beta_1 - \beta_3$ **0.84%**	

Table 2: Average increase in SPA solution relative to optimal solution

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%

Table 3: Test problems and results from second computational experiment

7. 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 system with multiple demand classes. We then discuss possible extensions to this research.

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 non-zero 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 twenty-four-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 non-zero. Let w_i be the service time for demand class i. We assume each $w_i < L$. 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 $c_{j-1} < IOH \le c_j$ for some $j \in \{|1, N|\}$. Then, if the next demand ***due*** were from class i, $i \in \{|1, j|\}$, it is served and the inventory level *IOH* is reduced by one; if the next demand ***due*** were from class i, $i \in \{|j+1, N|\}$, then it is backordered.

This policy is not optimal as it ignores information about 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.

With these assumptions, we can re-state the steady-state equations analogous to (5) and (6):

$$IL_N = IP_N - \sum_{i=1}^{N} D_i \left(L - w_i \right) \tag{14}$$

$$IL_i = s_i - B_{i+1,i} \quad for \ i = 1, 2, \cdots N \text{ -1} \tag{15}$$

where $D_i \left(\tau \right)$ 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 $\lambda_i \tau$. The equations (7), (8), (9) are the same.

As explanation, we refer to the operation of the N-stage serial inventory system. When a demand (either external or internal) occurs at time *t* at stage $i \in \left\{ \left| 1, N \right| \right\}$ with a due date of *t* + *w*, we assume that stage *i* does not fill this demand until its due date; if the stage cannot fill the demand on the due date, then the stage backorders the demand until it has inventory to fill it. We also assume that at time *t* stage $i \in \left\{ \left| 1, N-1 \right| \right\}$ initiates a one-for-one replenishment from its upstream supplier but with the due date of *t* + *w*. When a demand (either external or internal) occurs at time *t* at stage *N* with a due date of *t* + *w*, we assume that stage *N* reduces its inventory position by one; when its inventory position reaches a reorder-point equal to $s_N = R - c_{N-1}$, stage *N* orders on an external supplier.

The last step in developing the model for non-zero service times is to indicate the allocation scheme for filling backorders at each stage. We assume that at each stage we fill the backorders in the order of their due dates, with no differentiation between external and internal backorders. As a consequence, if $B_i \left(t \right) > 0$ for some stage *i*, then we can express the internal and external backorders as:

$$B_i \left(t \right) = \sum_{j=1}^{i} D_j \left(\tau_i - w_i, t - w_i \right)$$

where $\tau_i < t$ is the most recent time at which stage *i* stocks out. With this assumption we have that the probability distribution of internal backorders, conditioned on the total backorders at a stage, is binomial, as given by (10).

We thus see that the model (5) – (9) extends directly to permit non-zero 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.

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 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 the central warehouse replenishes its inventory with a one-for-one replenishment from an external supplier with a deterministic lead-time, and fills order from the local sites on a first-come, first-served basis with a deterministic lead-time. These assumptions are quite typical for low-volume, high-value service parts.

We can use the model (5) – (9) for each site, but with one modification. When the central warehouse stocks out, replenishment requests from the local sites are delayed. Thus, we need re-state equation (5) for the inventory at local site *k* as:

$$IL_{N,k} = IP_{N,k} - \sum_{i=1}^{N} D_{i,k}^{L} - B_{0,k} \qquad (16)$$

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 (16) and (6) – (9) to optimize the inventory parameters at each local site, as described in this paper.

General Demand Process. We assume a Poisson demand process. The model (1) – (4) remains valid for demand processes with independent increments, e.g., compound Poisson demand. However, we would need to re-visit the next steps in the model development if demand were from a compound Poisson process. The distribution of the inventory position in equation (5), 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 (10). 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., $D_i\left(s,t\right)$ is normally distributed with mean and standard deviation given by $\left(t-s\right)\mu_i$ and $\sqrt{t-s}\sigma_i$. As this process has independent increments, we can apply the model (1) – (4), 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 (10), 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.

Allocation Process. We assume that when the replenishment quantity Q is not sufficient to cover all backorders, we fill the backorders in the order of occurrence with no differentiation between external and internal backorders. The intent is to allocate the replenishment quantity fairly between filling the backorders at lower-ranked demand classes and restoring the reserve stock for higher-ranked demand classes. Nevertheless, this process is independent of any objective function, and is not optimal. It would be of interest to understand better how this allocation scheme performs for various problem criteria.

Lost Sales. The development of the model in this paper depends on the assumption that demand is backordered when it cannot be met from stock. We have not found an easy way to modify the current model to accommodate a lost sales assumption. We leave this for future research.

Cost Minimization Problem. In the text we develop our analysis and solution procedure for one specification of the inventory problem, namely the Service Level Problem. One might consider another problem specification whereby we minimize the total inventory-related costs, with no service-level constraints. The objective function would include the expected inventory holding costs, plus the expected backorder costs. For instance, suppose we have an inventory holding cost of *h* per unit per unit time, and a backorder cost for demand class *i* of b_i per unit per unit time, with $b_i \geq b_{i+1}, i \in \left\{ \left| 1, N-1 \right| \right\}$.

We suggest a single-pass algorithm for this problem, in which we solve a newsboy problem for each stage (demand class) in the serial inventory system, starting with stage *N*. For each class *i* the overage cost for the newsboy problem is $c_i^o = h$. We propose the following recursion for setting the underage cost:

$$c_1^u = b_1$$

$$c_i^u = \frac{\lambda_i b_i}{\sum_{j=1}^{i} \lambda_j} + \left(1 - \frac{\lambda_i}{\sum_{j=1}^{i} \lambda_j} \right) \left(1 - \beta_{i-1} \right) c_{i-1}^u$$

$$\text{where } \beta_i = \frac{c_i^u}{c_i^u + c_i^o}$$

As explanation, the underage cost needs to reflect the expected cost of both an external backorder as well as an internal backorder. The cost of an external backorder is the stage's backorder cost b_i. The cost of an internal backorder is the underage cost at stage *i-1* times the probability that stage *i-1* is out of stock; we estimate this probability from the fill-rate β_{i-1} imputed from the newsboy problem at stage *i-1*. We get the underage cost at stage *i* from a demand-weighted average of the cost of an external and internal backorder.

We expect this algorithm to perform similarly to the SPA for the Service-Level Problem. It should provide near-optimal solutions with a modest computational effort. Nevertheless, we leave it to future research to explore this.

References

Atkins, D. and Katircioglu, K. 1996. "Managing Inventory for Multiple Customers Requiring Different Levels of Service," Working Paper, University of British

Columbia, Vancouver, Canada, B.C.

Dekker, R., R.M. Hill, and M.J. Kleijn, 1997. "On the $(S-1, S)$ Lost Sales Inventory Model with Priority Demand Classes," Technical Report 9743/A, Econometric Institute, Erasmus University Rotterdam, The Netherlands.

Deshpande, V., M.A. Cohen, and K. Donohue, 2003. "A Threshold Inventory Rationing Policy for Service-Differentiated Demand Classes," *Management Science*, 49(6), 683-703.

Evans, R.V., 1968. "Sales and Restocking Policies in a Single Inventory System," *Management Science*, 14(7), 463-473.

Frank, K.C., R.Q. Zhang, and I. Duenyas, 2003. "Optimal Policies for Inventory Systems with Priority Demand Classes," *Operations Research*, 51(6), 993-1002.

Graves, S.C., 1985. "A Multi-Echelon Inventory Model for a Repairable Item with One-for-one Replenishment," *Management Science,* 31(10), 1247-1256.

Graves, S.C., 1996. "A Multi-Echelon Inventory Model with Fixed Replenishment Intervals," *Management Science,* 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 Science*, 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 Research Logistics*, 44(5), 457-472.

Kaplan, A., 1969. "Stock Rationing," *Management Science*, 15(5), 260-267.

Melchiors, P., R. Dekker, and M.J. Kleijn, 2000. "Inventory Rationing in an **(s, Q)** Inventory Model with Lost Sales and Two Demand Classes," *Journal of Operations Research Society*, 51(1), 111-122.

Moon, I. and S. Kang, 1998. "Rationing Policies for Some Inventory Systems," *Journal of the Operations Research Society*, 49(5), 509-518.

Nahmias, S. and S. Demmy, 1981. "Operating Characteristics of an Inventory System with Rationing," *Management Science*, 27(11), 1236-1245.

Topkis, D., 1968. "Optimal Ordering and Rationing Policies in a Non-stationary Dynamic Inventory Model with n Demand Classes," *Management Science*, 15(3), 160-176.

Veinott, A.F., 1965. "Optimal Policy in a Dynamic, Single Product, Nonstationary Inventory Model with Several Demand Classes," *Operations Research*, 13(5), 761-778.

de Véricourt, F., F. Karaesmen, Y. Dallery, 2000. “Dynamic Scheduling in a

Make-to-stock System: A Partial Characterization of Optimal Policies," *Operations Research*, 48(5), 811-819.

de Véricourt, F., F. Karaesmen, Y. Dallery, 2002. "Optimal Stock Allocation for a Capacitated Supply System," *Management Science*, 48(11), 1486-1501.

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