

A THRESHOLD INVENTORY RATIONING POLICY FOR SERVICE DIFFERENTIATED DEMAND CLASSES

Vinayak Deshpande

Krannert School of Management, Purdue University
W. Lafayette, IN 47907

Morris A. Cohen

Department of Operations and Information Management
The Wharton School of the University of Pennsylvania
Philadelphia, PA 19104-6366

Karen Donohue

Department of Operations and Management Science
Carlson School of Management, University of Minnesota
Minneapolis, MN 55455

July 2002

Abstract

Motivated by a study of the logistics systems used to manage consumable service parts for the U.S. military, we consider a static threshold-based rationing policy that is useful when pooling inventory across two demand classes characterized by different arrival rates and shortage (stockout and delay) costs. The scheme operates as a (Q, r) policy with the following feature. Demands from both classes are filled on a first-come-first-serve basis as long as on-hand inventory lies above a threshold level K . Once on-hand inventory falls below this level, low priority (i.e., low shortage cost) demand is backordered while high priority demand continues to be filled. We analyze this static policy first under the assumption that backorders are filled according to a special threshold clearing mechanism. Structural results for the key performance measures are established to enable an efficient solution algorithm for computing stock control and rationing parameters (i.e., Q, r , and K). Numerical results confirm that the solution under this special threshold clearing mechanism closely approximates that of the priority clearing policy. We next highlight conditions where our policy offers significant savings over traditional 'round-up' and 'separate stock' policies encountered in the military and elsewhere. Finally, we develop a lower bound on the cost of the optimal rationing policy. Numerical results show that the performance gap between our static threshold policy and the optimal policy is small in environments typical of the military and high technology industries.

This research was supported in part by the U.S. Navy under Contract # NOV391-96-M-M04, NSF CAREER Award #9602072, NSF grant #0075391, and the Fishman-Davidson Center for Service Operations Management at the Wharton School. Special thanks to the following people from various organizations within the military - Sandy Leggieri, Gary Burchill, Jere Engelman, and Mike Puoy. The authors also acknowledge the detailed and insightful comments of Bill Lovejoy, the AE and two anonymous referees on earlier versions of this paper.

1 Introduction

The practice of rationing inventory (or capacity) among different customer classes is an increasingly important tool for balancing supply with demand in environments where requirements for service vary widely. The practice of issuing stock to some customers while refusing or delaying demand fulfillment for others is analogous to the highly successful yield management policies adopted by airlines and hotels in recent years. In this paper, we analyze a stock rationing scheme that is useful for managing inventory in a continuous review (Q, r) environment with two customer demand classes defined by unique arrival rates and service costs. The scheme is characterized by a threshold inventory level, K , which signals when to reserve stock for higher priority customers. The associated (Q, r, K) inventory policy serves all customers on a first-come-first-serve basis while on-hand inventory is above K , and cuts off service to low priority customers when on-hand inventory fall below this threshold.

Our interest in this policy grew from an empirical study of the military’s logistic system supporting service parts for military weapon systems (Cohen et al. 1998). The military recently moved the management of these parts from the individual military services (e.g., separate Army and Navy warehouses) to a central inventory control point within the Defense Logistics Agency (DLA). While this change offers inventory pooling benefits for common parts, it has led to some disagreement across the military services about the appropriate safety stock levels. The disagreement stems from the fact that the criticality of a part often differs significantly for each military service. DLA’s current policy for managing these demand classes is to “round-up” each part’s availability requirements across the various military services. For example, if the Army requires a service level of 85% while the Navy requires 95%, DLA stocks the part to meet an aggregate service level of 95%. Once stocked, inventory is allocated to customers on a first-come-first-serve (FCFS) basis. There are two obvious shortcomings to this approach. First, by rounding-up requirements, DLA may be investing too much inventory in non-critical items. Second, processing orders on a FCFS basis allows a low priority customer to possibly preempt more critical customers. The military’s previous strategy of managing separate pools of stock for each service avoided these problems, but did so at the cost of no inventory pooling. A threshold rationing policy, similar to the (Q, r, K) policy studied here, has been proposed as a way to avoid the problems inherent in the round-up policy while still taking advantage of inventory pooling.

While our problem is motivated by the dynamics observed in the service parts division of the US military, we expect our solution approach is applicable to a wide range of industry settings.

Inventory systems with multiple demand classes having different priorities are common to a number of industries. For example, Cohen, Kleindorfer and Lee (1998) study a service parts application in the computer industry where a retailer could place normal replenishment orders and emergency orders, in case of stockout, at the warehouse. Kleijn and Dekker (1998) provide an overview of inventory systems with several demand classes, including examples ranging from airlines to petrochemical companies.

While such rationing policies have been implemented from time to time in the military, there is no methodology for determining how to select the parameters for these policies. The goal of this paper is to develop a tractable and implementable solution to the stock rationing problem, and offer managerial insights on conditions when our proposed policy is attractive. We do so by developing a methodology for selecting optimal control parameters for the (Q, r, K) policy, i.e. to select policy parameters to minimize inventory, delay and backordering costs. We also analyze the characteristics of our solution to provide insight into when threshold rationing offers significant benefits over traditional ‘round-up’ and ‘separate stock’ mechanisms. Finally, we explore the performance of our solution relative to optimal policies for stock rationing which may include non-stationary, state dependent allocation decisions of the non-threshold type.

Very little research exists on how to optimize policy parameters for multiple customer classes, particularly in cases with fixed setup costs, positive lead-times, and backlogged customer demand (Kleijn and Dekker, 1998). This environment is challenging because positive backorders and positive on-hand inventory can coexist at the same point in time, making it difficult to calculate backorder distributions from the inventory level distribution. We solve the optimization problem by first studying how the policy performs under a special backlog clearing mechanism which allows closed form expressions for the stockout levels, and average number of demands in backlog, for each demand class. Based on these results, we then develop an efficient algorithm for calculating the optimal control parameters (Q, r, K) for this environment and show numerically that the resulting optimal solution closely approximates the solution under a more preferred backlog clearing mechanism. Numerical results also reveal that, compared to the DLA’s current round-up policy, the (Q, r, K) policy is most beneficial when the arrival rate for low criticality demand is significantly higher than that of the higher criticality class. Using a round-up policy in this case is wasteful since a large amount of inventory is used to support the higher service level for the low criticality demand class.

While the (Q, r, K) policy proposed here is easy to implement and performs well compared to

traditional policies, other non-stationary policies (i.e., where K may vary with the state of the system) could perform better. A secondary goal of this paper is to establish when the threshold (Q, r, K) policy is a reasonable approximation to the optimal non-stationary rationing policy. This is accomplished by developing a lower bound over all possible policies. Numerical results suggest that the performance gap between our static threshold rationing policy and the optimal non-stationary rationing policy is small for cost and demand parameters typical of military service parts and other environments where setup costs are extremely high (e.g., semiconductor equipment which is a high technology, make to order, capital intensive industry). It does not perform as well when both setup costs are small and penalty costs for the two demand class are significantly different.

The paper continues in section 2 where we position our research with respect to previous literature. Section 3 introduces the threshold rationing model, backlog clearing mechanisms, and cost function that drives our methodology. Performance measures, structural results, and a solution algorithm are developed in section 4 assuming a (Q, r, K) policy operating with a special backlog clearing mechanism called “threshold clearing”. Section 5 compares these results with a more attractive “priority clearing” mechanism. Section 6 provides insight into the benefits of threshold rationing over DLA’s current approaches, and compares our results to a lower bound on the cost of an optimal non-stationary rationing policy. We also test the assumption of independent Poisson demand processes by comparing our results to a perfectly correlated demand case. Section 7 provides a formulation of our rationing policy for more than 2 demand classes and briefly discusses how the analysis would change to accommodate this more complex system. The paper concludes in section 8 with a discussion of possible extensions.

2 Literature Review

The task of dynamically allocating inventory to different demand classes lies at the heart of many yield management problems. These problems are typically characterized by limited capacity and perishable inventory (e.g., seats on an airplane, cars in a rental fleet, or rooms in a hotel) which is allocated to different classes of demand (e.g., first class, business class, or economy). Kimes (1989) provides an overview of research in this area. In this environment, the key decision variables are normally the *prices* charged to each demand class as well as the possible *rationing levels* (i.e., booking limits) to impose. Some examples include Belobaba (1989) who examines booking limits for airline seats with different price classes, and Bitran and Gilbert (1996) who develop heuristic

rationing procedures for managing hotel reservations. In these problems, the capacity or inventory level is fixed so the decision of how much inventory to order and when to replenish are not relevant. The presence of obsolescence, however, leads to non-stationary control policies that dynamically adjust as time to expiration approaches. Examples of dynamic allocation models for yield management include Lee and Hersh (1993), Bitran and Mondschein (1995), Subramanian et al. (1998), and Zhao and Zheng (2001). The major differences between our stock rationing problem and traditional yield management problems are that we allow for multiple replenishment opportunities and assume inventory is not perishable. We also focus on static, rather than dynamic, policies.

Turning to the inventory literature, Veinott (1965) was one of the first to consider multiple demand classes in a multi-period, single product, non-stationary inventory environment. While he focuses on the question of how much to order and when to replenish, he does so in the context of a periodic review system without rationing levels. Topkis (1968) extends Veinott's work by considering how inventory should be allocated between demand classes within a single period of a periodic review model. Here each demand class is characterized by a different shortage cost. The analysis is facilitated by breaking each review interval into a finite number of sub-periods. At the end of each sub-period, the decision maker allocates inventory to demand that has been realized thus far. The allocation is based on a trade-off between the benefit of filling demand for low class items in the current sub-period and reserving inventory to fill higher class items in subsequent sub-periods. Within a single review interval, Topkis proves there exists optimal, non-negative, rationing levels for each demand class which, under certain conditions, are decreasing in time.

Our rationing policy differs from Topkis' in three fundamental ways. First, we make the decision of whether to fill or delay an order at the moment the order arrives. Topkis delays this decision until the end of each sub-period. Making the decision up front reduces order delays. Second, our rationing level is stationary, which is consistent with our continuous review environment, where there are no defined time intervals for revising decisions. Third, our replenishment order cycles are based on inventory position, taking into account setup costs, lead-times, and the possibility that multiple replenishment orders may be in the pipeline. Models similar to Topkis under different operating environments have been considered by Kaplan (1969) and Frank et al. (1999).

Nahmias and Demmy (1981) were the first to analyze a rationing policy in a (Q, r) environment. They consider a continuous review system with Poisson demand and two demand classes (as we do). However, they focus on evaluating fillrates for given rationing and reorder levels rather than on optimizing the policy parameters (Q, r, K) under a cost framework. Our model formulation also

differs from theirs in that we do not require their simplifying assumption that not more than one order is outstanding at any point in time.

Ha (1997a, 1997b) considers a similar rationing policy in the context of a single-item, make-to-stock, production system with two or more demand classes. Assuming Poisson demand and exponential production times, the optimal policy is characterized by a sequence of monotone and stationary rationing levels. Recently, Ha (2000) extended this analysis to the case of Erlang distributed processing time. Vericourt et al. (2000, 2002) also recently developed a characterization of the optimal policy for the backorders case with zero set-up costs and exponential lead-times.

Another example of the stock rationing problem is the allocation of a common component to multiple products in an assemble-to-order system. Baker et al. (1986), Gerchak et al. (1988) and Gerchak and Henig (1986) all consider optimal ordering and rationing policies for a common component in an assemble-to-order environment. Most of these papers consider a single period model with multiple end products having both a common and a product specific component. The objective is to determine initial stocking levels for the common and product specific components, and to determine a rationing policy for the common component *after* realization of demand, to minimize inventory holding costs subject to fill-rate constraints. The end product fillrates are determined by the availability of both the common and product specific component. In our paper, there is a one to one correspondence between end product fillrates and common component availability. Our analysis can therefore be considered an extension of the component commonality literature to situations with a continuous review infinite horizon, setup costs and positive lead-times, but with no product specific components.

The queueing literature also considers the impact of admission control policies (e.g., first-come-first-serve, earliest due-date, or highest delay cost) for multiple customer classes. Ross and Tsang (1989), for example, develop a stochastic knapsack model for allocation of servers to arriving customers in a model relevant to telecommunications networks and rental car fleet management. Savin et al. (2000) provide an analysis of a multi-class environment in the rental business. It is interesting that threshold-like rationing policies are used here as well. In these problems, capacity is fixed and stock returns after a random service time. In our environment, demand continuously depletes stock, while stock is replenished by placing replenishment orders.

Besides adopting a rationing policy, other researchers have considered managing multiple demand classes through various priority mechanisms. Cohen, Kleindorfer and Lee (1988) use a simple priority mechanism to allocate stock in a multi-echelon inventory system. They assume a (s, S)

policy with two demand classes and consider two replenishment modes, emergency and normal, with different lead-time lengths. An aggregate fill-rate constraint based on the aggregate demand for the two classes is imposed as opposed to individual fill-rate constraints (individual stockout costs in our model) and they do not ration inventory between the two demand classes. Our model assumptions, in general, are quite different. For example, in keeping with DLA’s current ordering environment, we assume a continuous review (Q, r) inventory policy is followed with a constant replenishment lead-time.¹

3 Model Framework

In keeping with the military environment that motivated this work,² we assume inventory for an item is held and replenished over time to fill reoccurring demand from two customer classes $i = 1, 2$. Section 7 provides insight into how the problem formulation and solution would change for three or more classes. We assume demand from class i follows a Poisson process with rate λ_i , implying a total demand rate of $\lambda = \lambda_1 + \lambda_2$. Any unmet demand is backlogged and incurs two penalty costs: a stockout cost per unit backordered (π_i) and a delay cost per unit per period of delay ($\hat{\pi}_i$), where $i = 1, 2$. With no loss of generality we assume $\pi_1 \geq \pi_2$ and $\hat{\pi}_1 \geq \hat{\pi}_2$, and therefore refer to class 1 demand as having ‘higher priority’.

Inventory is replenished according to a (Q, r, K) policy that operates as follows. When the inventory position (on-hand plus on-order minus backorders) reaches the level r , a replenishment order for Q units is placed and arrives $\tau > 0$ time units later. Demands from both classes are filled on a FCFS basis as long as the on-hand inventory level is greater than or equal to K . Once the on-hand inventory level falls below K , class 2 demand is backlogged (i.e., no longer filled) while class 1 demand continues to be filled as long as inventory is available.

Figure 1 illustrates a typical inventory cycle for the (Q, r, K) policy. In this example, K is set lower than the reorder point r , although this is not required in general. Here on-hand inventory initially depletes at the aggregate demand rate λ . Once on-hand inventory falls to K , the depletion rate reduces to λ_1 since class 2 demands are now backlogged. Notice that class 2 backorders may exist when there is positive on-hand inventory, while class 1 backorders only occur when the system

¹DLA reviews its inventory position every two to three days and places an order for an economic lot whenever the position falls below the reorder level. Given the extremely low demand rates for many SKUs in this environment (sometimes less than five per year) a continuous review approximation is quite appropriate. Also, the typical life cycle of a weapon system lasts about fifteen years, with the post-introduction phase lasting more than ten years. Thus a stationary (Q, r) model framework is appropriate for parts in the stable operational phase of their lifecycle.

²While DLA experiences as many as twelve demand classes, the percentage of parts shared across more than two demand classes is relatively small. Thus the two class case is thought to capture most of the pooled demand.

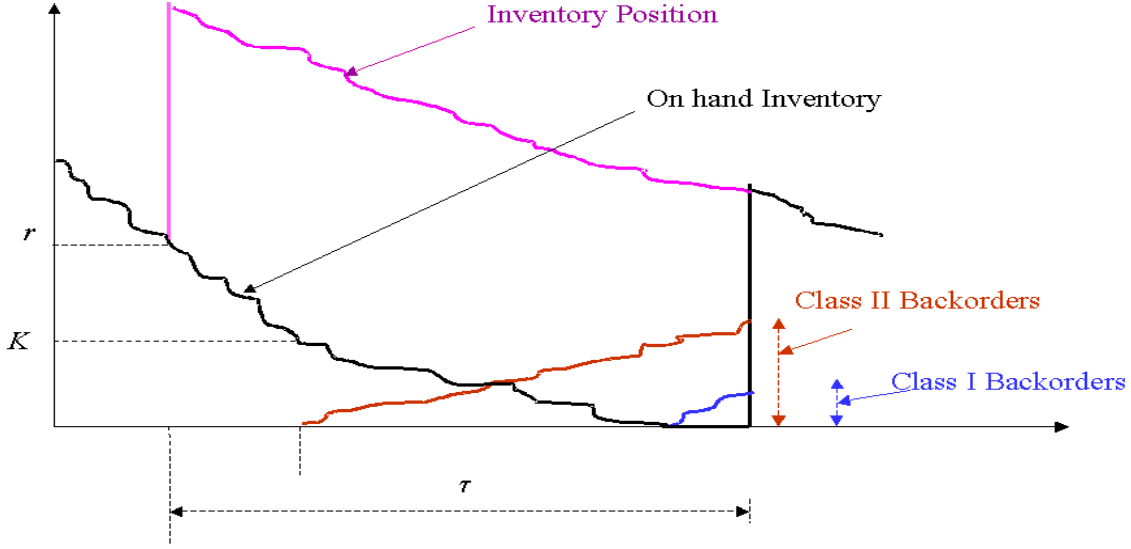


Figure 1: Typical Cycle for a (Q, r, K) Policy

runs out of stock.

Our objective is to determine the policy parameters (Q, r, K) which minimize expected annual cost for the system. We assume that each replenishment order incurs a fixed setup cost of s , while inventory holding costs are incurred at rate h for each unit of inventory carried on-hand. Let $C(Q, r, K)$ denote the expected annual cost for a given (Q, r, K) policy and $H(Q, r, K)$, $S(Q, r, K)$, and $Z(Q, r, K)$ denote the associated expected annual holding, setup, and penalty costs, respectively. Our problem is then stated as

$$\min_{Q, r, K, K \leq r+Q} C(Q, r, K) \quad (1)$$

$$C(Q, r, K) = S(Q, r, K) + H(Q, r, K) + Z(Q, r, K).$$

Note that since the maximum possible on-hand inventory is $r + Q$, we limit our search for optimal threshold rationing level to $K \leq r + Q$.

Before proceeding further, it is important to position our proposed (Q, r, K) policy relative to the larger family of possible rationing policies for two class systems. In general, a rationing policy describes when to fill orders from a particular customer class. In our setting, a rationing policy tries to balance the smaller but certain class 2 penalty cost (incurred when holding back a current class 2 customer) against the possibility of a greater but uncertain class 1 penalty cost. All rationing policies provide guidance on when to hold back inventory from lower priority (i.e., class 2) customers. This guidance can take the form of a static threshold level, in the case of our (Q, r, K) policy, or a non-stationary threshold level, which changes based on the state of the system

(e.g., time until next replenishment). Intuitively, the optimal non-stationary policy can dominate the static policy. In section 6, we quantify how large the cost gap can be, by computing a lower bound on the cost of the optimal rationing policy.

Static threshold rationing policies can be further characterized by the mechanism they use to clear backlog orders when a replenishment order arrives. Note that Nahmias and Demmy (1981) ignore this issue by assuming at most one order is outstanding at any point in time. The most obvious clearing mechanism is simple “priority clearing”, which gives priority to class 1 backorders and only fills class 2 backorders if on-hand inventory (after filling all class 1 backorders) is greater than K . Clearing the backlog in this fashion will minimize cost since class 1 stockout costs are always greater than class 2 stockout costs. Unfortunately, this priority clearing mechanism is difficult to analyze analytically since the on-hand inventory and backorders depends in a complicated way on the order arrival process. For this reason, we introduce an alternative backorder clearing mechanism, called “threshold clearing”, which serves as an approximation to the “priority clearing” mechanism. The idea of this mechanism is to clear backorders in the same manner as orders would be filled had there been more inventory available at the time demand arrived. Although this mechanism may clear some class 2 backorders before class 1, the probability of that happening will be quite low if the fillrate for class 1 demand is reasonably high.

In the next section, we more formally introduce this “threshold clearing” mechanism and focus on solving the following problem:

$$\begin{aligned} \min_{Q,r,K,K \leq r+Q} \quad & C^T(Q,r,K), \\ \text{where} \quad & C^T(Q,r,K) = S^T(Q,r,K) + H^T(Q,r,K) + Z^T(Q,r,K) \end{aligned} \tag{2}$$

and the superscript T denotes the associated “threshold clearing” mechanism. We show in section 5 that the solution to this problem closely approximates the solution to problem (1) (i.e., a (Q,r,K) policy using the priority clearing mechanism) for a wide range of problem parameters.

4 Analysis of the (Q,r,K) policy under a threshold clearing mechanism

Our purpose in this section is to develop expressions for the key performance measures needed to evaluate $C^T(Q,r,K)$, provide structural properties for these measures, and use these structural properties to define an efficient algorithm for solving problem (2). Note that both problems (1) and (2) require developing expressions for the limiting on-hand inventory distribution and limiting class 1 and 2 backorder distributions. It is well known that the inventory position process $IP(t)$ and the

inventory level process $IL(t)$ have limiting distributions (see Hadley and Whitin 1963, and Zipkin 1986a). Let $IP(\infty)$ and $IL(\infty)$ denote the random variables with these limiting distributions, then

$$IL(\infty) = IP(\infty) - LD(\infty)$$

where $IP(\infty)$ is uniformly distributed on $\{r + 1, r + 2, \dots, r + Q\}$ and $LD(\infty)$ represents total lead-time demand which has a Poisson distribution with mean $\mu = \lambda\tau$. In a traditional (Q, r) policy, it is easy to compute the on-hand inventory and backorder distributions from the limiting inventory level distribution, because on-hand inventory is the positive component of inventory level while backorders is the negative component of inventory level. However for a (Q, r, K) policy, the inventory level is the on-hand inventory net of *all* backorders. In other words,

$$IL(t) = OH(t) - BO_1(t) - BO_2(t),$$

where $OH(t)$ denotes on-hand inventory and $BO_i(t)$ denotes class i backorders, $i = 1, 2$, all at time t . Due to rationing of class 2 demand, class 2 backorders and on-hand inventory can be non-zero *simultaneously* for a (Q, r, K) policy. Hence inventory level alone does not provide sufficient information to characterize on-hand inventory and backorder levels. In fact, these levels depend not only on how inventory is rationed but also on how backorders are cleared on arrival of a replenishment order. The “threshold clearing” mechanism, alluded to in the definition of problem (2), allows us to compute these levels with a minimal amount of state information.

4.1 The threshold clearing mechanism

Before proceeding with the analysis of problem (2), we need to explain more formally how the threshold clearing mechanism works. Note first that clearing mechanisms only come into play when backorders exist on arrival of a replenishment order. In this case, backorders may have grown so large that they cannot all be accommodated by a replenishment order without dropping the on-hand inventory level below K . The idea of the threshold clearing mechanism is to clear backorders as if the on-hand inventory was $r + Q$ a lead-time back and the threshold rationing policy was followed subsequently.

Figure 2 shows the sequence of events during a typical cycle. Here the j^{th} replenishment order is placed at time t_j , causing the inventory position to rise to $r + Q$. Time t_{B_2} marks the point when on-hand inventory first hits K and class 2 demands begin to be backlogged. Time t_{B_1} marks the point when on-hand inventory falls to zero and class 1 demand also begins to backlog. We also define time t_{K_j} as the time of the $r + Q - K^{th}$ demand arrival in the interval $(t_j, t_j + \tau)$.

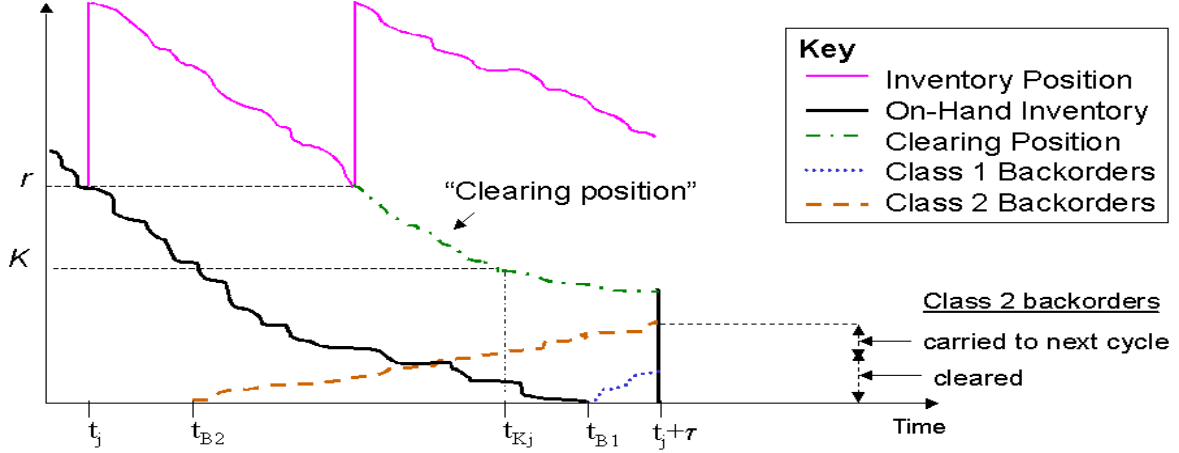


Figure 2: The Threshold Clearing Mechanism

The “clearing position” at time $t \leq t_{K_j}$ reflects the inventory position at time t_j (which is $r + Q$) minus all subsequent demand during the interval (t_j, t) . The “clearing position” at time $t > t_{K_j}$ reflects the clearing position at time t_{K_j} (which is K) minus subsequent class 1 demand during the interval $(t_{K_j}, t_j + \tau)$. The threshold clearing mechanism uses this critical time t_{K_j} to separate which backorders to clear once the j^{th} replenishment order arrives at time $t_j + \tau$. The general rules are

1. clear all (class 1 and class 2) backlogged demand that arrived before t_{K_j} in the order of arrival (FCFS).
2. clear any remaining backlogged class 1 demand until either all class 1 backorders are filled or no on-hand inventory remains.
3. carry over (i.e., continue to backlog) all class 2 demand arriving after t_{K_j} .

This rule effectively allocates clearing inventory above K to both customer classes while reserving any remaining inventory for class 1 backorders. This may result in more class 1 backorders than using a priority clearing mechanism, which clears all class 1 backorders first.

The beauty of this clearing mechanism is that it allows us to calculate important performance measures with limited state information. For example, to calculate the number of class 1 and 2 backorders carried over for a given replenishment period, let $D_i(t_j, t_j + \tau)$ denote the number of class i demands that arrive between the placement and receipt of replenishment order j , where $D(t_j, t_j + \tau) = D_1(t_j, t_j + \tau) + D_2(t_j, t_j + \tau)$. When a new replenishment of size Q arrives, at time $j + \tau$, one of two things can happen. If $D(t_j, t_j + \tau) \leq (r + Q - K)$, then we can clear all backorders

and raise on-hand inventory to $r + Q - D(t_j, t_j + \tau)$. Otherwise, some backlog may be carried over until the arrival of the next replenishment order. The on-hand inventory after replenishment order j arrives and the associated backlog is cleared is then

$$OH(t_j + \tau) = \begin{cases} r + Q - D(t_j, t_j + \tau) & \text{if } (r + Q - K) \geq D(t_j, t_j + \tau) \\ [K - D_1(t_{K_j}, t_j + \tau)]^+ & \text{otherwise,} \end{cases} \quad (3)$$

and the number of backorders remaining for class 1 and class 2 demand is

$$BO_1(t_j + \tau) = [D_1(t_{K_j}, t_j + \tau) - K]^+, \quad (4)$$

$$BO_2(t_j + \tau) = D_2(t_{K_j}, t_j + \tau). \quad (5)$$

These performance measures only require knowledge of the inventory position information at time t_j and the demand arrivals in the interval $(t_j, t_j + \tau)$.

4.2 Performance Measures

We now develop expressions for the key performance measures underlying the cost function in problem (2). We first calculate the limiting on-hand inventory distribution and limiting class 1 and class 2 backorder distributions. We then calculate the long run fraction of time the system is out of stock and the average backorders for both demand classes.

The threshold clearing mechanism defined in the earlier sub-section helps us compute the performance measures from the inventory position distribution and the lead-time demand distribution. Note that equation (3) defines on-hand inventory at a specific replenishment order time, $t_j + \tau$. Equations (3-5) can be generalized to any random time t as follows.

$$OH(t + \tau) = \begin{cases} y - D(t, t + \tau) & \text{if } (y - K) \geq D(t, t + \tau) \\ [K - D_1(t_K, t + \tau)]^+ & \text{otherwise,} \end{cases} \quad (6)$$

$$BO_1(t + \tau) = [D_1(t_K, t + \tau) - K]^+, \quad (7)$$

$$BO_2(t + \tau) = D_2(t_K, t + \tau) \quad (8)$$

where, y is the inventory position ($IP(t)$) at time t , and t_K is now defined as the time of $y - K^{th}$ demand arrival in the interval $(t, t + \tau)$. Also, if $y < K$ then t_K is defined as the last-time inventory position hit K before reaching y at time t .

The above equations enable us to compute the steady-state on-hand inventory distribution, and the backorder distributions for the two classes. To compute the steady-state on-hand inventory distribution, it is useful to look separately at on-hand inventory distribution above K and below K .

We first focus on the on-hand inventory distribution above K . From equation (6), we know that $OH(t + \tau) \geq K$ whenever $IP(t) - D(t, t + \tau) \geq K$. Let j represent a possible value for $OH(t + \tau)$ and y denote the inventory position at time t . Conditioning on y , the probability that on-hand inventory equals j , for any $j \geq K$, is simply

$$Prob[OH(t + \tau) = j | IP(t) = y, j \geq K] = p(y - j; \lambda\tau) \quad \text{if } y \geq j > 0 \quad (9)$$

where $p(y - j; \lambda\tau)$ denotes the Poisson probability mass function of the demand process over the replenishment lead-time τ .

Now consider the on-hand inventory distribution below K . This case is more complicated since we now need to keep track of class 1 demand arrivals in the interval $(t_K, t + \tau)$ in equation (6). Let $\alpha_i = \frac{\lambda_i}{\lambda}$ denote the probability of an arrival being class $i, i = 1, 2$. The probability that we have exactly n_i class i demands from a demand stream of n customers is then a simple binomial $b(\alpha_i; n; n_i) = \frac{n!}{n_i!(n - n_i)!} \alpha_i^{n_i} (1 - \alpha_i)^{n - n_i}$. Hence, the probability that $D_1(t_K, t + \tau) = z$ is equal to $b(\alpha_1; x - y + K; z)p(x; \lambda\tau)$, where x represents the total number of demand arrivals in $(t, t + \tau)$, and $x - y + K$ represents the total number of demand arrivals in the interval $(t_K, t + \tau)$. Hence, conditioning on y , the probability that on-hand inventory equals j , for any $j < K$, is then

$$Prob[OH(t + \tau) = j | IP(t) = y, j < K] = \begin{cases} \sum_{x=y-j}^{\infty} b(\alpha_1; x - y + K; K - j)p(x; \lambda\tau) & \text{if } 0 < j \leq y, \\ \sum_{x=y}^{\infty} \sum_{z=K}^{x-y+K} b(\alpha_1; x - y + K; z)p(x; \lambda\tau) & \text{if } j = 0, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Combining cases (9) and (10) and taking limits, we have

$$Prob[OH(\infty) = j | IP(\infty) = y] = \begin{cases} p(y - j; \lambda\tau), & \text{if } y \geq j \geq K, j > 0, \\ \sum_{x=y-j}^{\infty} b(\alpha_1; x - y + K; K - j)p(x; \lambda\tau) & \text{if } 0 < j < K, \\ \sum_{x=y}^{\infty} \sum_{z=K}^{x-y+K} b(\alpha_1; x - y + K; z)p(x; \lambda\tau) & \text{if } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

We can now obtain the limiting on-hand inventory distribution by unconditioning over the limiting inventory position y , which is uniformly distributed over $(r + 1, r + Q)$. This gives

$$Prob[OH(\infty) = j] = \frac{1}{Q} \sum_{y=r+1}^{r+Q} Prob[OH(\infty) = j | IP(\infty) = y] \quad (11)$$

Using equations (7-8), this same conditioning logic can be used to derive expressions for the limiting backorder distributions for each demand class $i, i = 1, 2$.

$$Prob[BO_i(\infty) = j] = \frac{1}{Q} \sum_{y=r+1}^{r+Q} Prob[BO_i(\infty) = j | IP(\infty) = y], \quad (12)$$

where

$$Prob[BO_1(\infty) = j | IP(\infty) = y] = \begin{cases} \sum_{x=y+j}^{\infty} b(\alpha_1; x - y + K; K + j)p(x; \lambda\tau) & \text{if } j > 0 \\ 1 - (\sum_{h=1}^{\infty} \sum_{x=y+h}^{\infty} b(\alpha_1; x - y + K; K + h)p(x; \lambda\tau)) & \text{if } j = 0 \end{cases}$$

and

$$Prob[BO_2(\infty) = j | IP(\infty) = y] = \sum_{x=(j+y-K)^+}^{\infty} b(\alpha_2; x - y + K; j)p(x; \lambda\tau).$$

Note that for the class 2 backorder distribution, we count the fraction of class 2 arrivals in the $x - y + K$ arrivals during $(t_K, t_K + \tau)$ (rather than the fraction of class 1 arrivals) and therefore use the probability α_2 for the binomial distribution.

These limiting distributions allow us to compute a wide range of performance measures. For example, let $A_i(Q, r, K)$ denote the long run fraction of time the system is out of stock for class i demand. Then $A_1(Q, r, K)$ is simply the probability that on-hand inventory is zero in steady state, while $A_2(Q, r, K)$ is the probability that on-hand inventory is less than or equal to K . This can be easily obtained from the on-hand inventory distribution given by equation (11). After algebraic simplification, we have

$$A_i(Q, r, K) = \frac{1}{Q} \sum_{y=r+1}^{r+Q} a_i(y, K) \quad (13)$$

where

$$a_1(y, K) = \sum_{x=y}^{\infty} \sum_{j=0}^{x-y} b(\alpha_1; x - y + K; K + j)p(x; \lambda\tau)$$

and

$$a_2(y, K) = \begin{cases} \sum_{x=y-K}^{\infty} p(x; \lambda\tau) & \text{if } K \leq y \\ 1 & \text{if } K > y \end{cases}$$

Expression (13) is useful for computing the offshelf fill-rate of each demand class. Let $F_i(Q, r, K)$ denote the off-shelf class i fill-rate for our proposed (Q, r, K) policy. Using the PASTA property for Poisson arrivals, this fill-rate is simply

$$F_i(Q, r, K) = 1 - A_i(Q, r, K). \quad (14)$$

It is interesting to compare this fill-rate calculation with the approximation provided by Nahmias and Demmy (1981). Recall that their (Q, r, K) policy model assumes that *at most* one order may be outstanding at any point in time. This implies that when an order arrives, the order quantity always raises the on-hand inventory level above r and wipes out all demand backlog. We conducted a numerical study to determine when this assumption is a reasonable approximation (see Deshpande 2000 for details). As expected, their approximation is quite accurate when Q is large with respect to total lead-time demand. However, their approximation deteriorates as Q decreases and goes so

far as to compute negative fill-rates in extreme cases. The approximation for the class 2 fill-rate also deteriorates as the value of the threshold level approaches the reorder level r . For these parameter ranges, equation (14) offers a significant improvement over the previous approximation.

The long-run average number of backorders for each demand class can be obtained by taking the expectation of backorders using the backorder distribution defined in equation (12). Let $B_i(Q, r, K)$ denote this quantity for class i . We then have

$$B_i(Q, r, K) = \frac{1}{Q} \sum_{y=r+1}^{r+Q} b_i(y, K) \quad (15)$$

where

$$b_1(y, K) = \sum_{x=y}^{\infty} \sum_{j=0}^{x-y} j b(\alpha_1; x - y + K; K + j) p(x; \lambda\tau)$$

and

$$b_2(y, K) = \begin{cases} \sum_{x=y-k}^{\infty} \alpha_2(x - y + K) p(x; \lambda\tau) & \text{if } K \leq y \\ \lambda_2\tau + \alpha_2(K - y) & \text{if } K > y. \end{cases}$$

4.3 Structural Results

We now provide structural results on the convexity properties of our key performance measures with respect to K and r . Besides being interesting in their own right, these results are a necessary precursor to the development of an efficient solution algorithm for problem (2). In this section, we use the terms increasing and decreasing to mean non-decreasing and non-increasing respectively. Proofs of all formal results are provided in the Appendix.

Focusing first on the long run probability of shortage, we have

Lemma 1 *The long run probability of shortage for class 2 customers, $A_2(Q, r, K)$, is decreasing in r and increasing in K .*

Lemma 2 *The long run probability of shortage for class 1 customers, $A_1(Q, r, K)$, is decreasing in r and decreasing in K .*

Because an increase in the long run probability of storage implies an increase in fill-rate, these Lemmas confirm that increasing K improves the fill-rate of class 1 customers while lowering that of class 2 customers. On the other hand, increasing r has a positive impact on both class 1 and class 2 fill-rates, as one would expect from any (Q, r) type policy. These results are derived by confirming the appropriate sign of the first differences.

To gain some insight into the implications of Lemmas 1 and 2, Figure 3 illustrates the joint impact of the threshold level and the reorder level on fillrates for the case of two demand classes

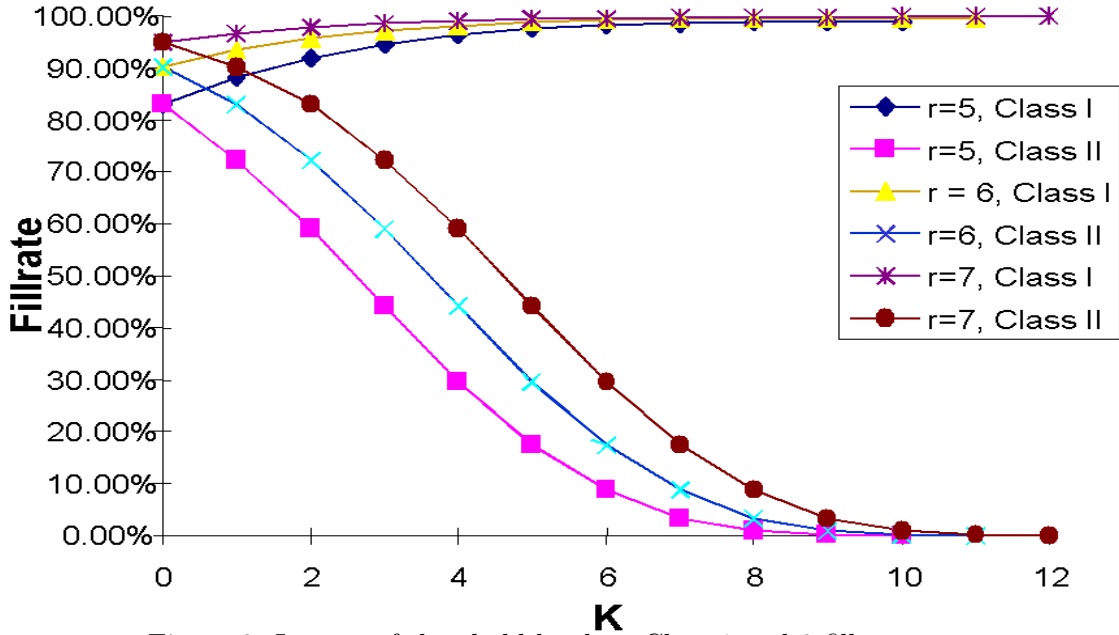


Figure 3: Impact of threshold level on Class 1 and 2 fill-rates

with average annual demand of 10 units and $Q = 5$. Fill rates for both classes increase with r , as implied by Lemmas 1 and 2. Also, the fill-rate for class 1 customers increases with K at the expense of the class 2 customers. It is interesting to note that the rate of deterioration in the class 2 fill-rate is severe as K increases. Fill-rates appear to be quite sensitive to the rationing parameter. Figure 3 also verifies that fill-rates for the two classes are equal when there is no rationing ($K = 0$). Moving toward a desired level of differentiation requires careful selection of both the threshold and reorder levels.

The next set of Lemmas show that the impact of K and r on the average backorder quantities is similar to what we just observed for the long run probability of shortage.

Lemma 3 *The average backorder quantity for class 2 customers, $B_2(Q, r, K)$, is decreasing in r and increasing in K .*

Lemma 4 *The average backorder quantity for class 1 customers, $B_1(Q, r, K)$, is decreasing in r and decreasing in K .*

While the first order effects of K and r are the same for A_i and B_i , there is an important difference in their rate of change. B_i is relatively well-behaved with respect to K and its inventory position y . In particular, B_i is decreasing convex in y . This is implied by the following Lemma.

Lemma 5 *$b_i(y, K)$ is convex in y for fixed K and convex in K for fixed y .*

This extends the results of Zipkin (1986b) and Zhang (1998) who prove the convexity of service level measures in a traditional (Q, r) model. In contrast, A_i is not convex in its parameters, although it does have a unique inflection point. The first difference for A_i initially decreases and then increases with y . To summarize, increasing the reorder level decreases both class 1 and class 2 delay costs but at a decreasing rate. However, increasing the reorder level decreases the stockout costs, initially at an increasing rate and then at a decreasing rate. We use these properties in the next section to establish the unimodality of our cost function.

4.4 Optimization Model and Solution Algorithm

We now formulate the terms of our cost function, $C^T(Q, r, K)$. Recall that total cost consists of setup, holding, and penalty costs. The average setup cost is simply the holding cost rate times the average inventory,

$$S^T(Q, r, K) = S(Q, r, K) = \frac{s\lambda}{Q}, \quad (16)$$

which is independent of the clearing mechanism used. The average holding cost is the unit holding cost times the average inventory,

$$H^T(Q, r, K) = h \sum_{j=0}^{r+Q} j \text{Prob}[OH(\infty) = j]. \quad (17)$$

The average on-hand inventory is computed using the on-hand inventory distribution given in equation (11). This expression can be rewritten as

$$H^T(Q, r, K) = h \left\{ \frac{(Q+1)}{2} + r - \mu + B_1(Q, r, K) + B_2(Q, r, K) \right\} \quad (18)$$

where μ denotes mean lead-time demand (i.e., $\lambda\tau$). Finally, the total average penalty cost (i.e., cost of stockouts and delays) is

$$Z^T(Q, r, K) = \sum_{i=1}^2 \pi_i \lambda_i A_i(Q, r, K) + \sum_{i=1}^2 \hat{\pi}_i B_i(Q, r, K). \quad (19)$$

Plugging expressions (13) and (15) into equations (16)-(19) and simplifying, our objective function becomes

$$C^T(Q, r, K) = \frac{s\lambda + \sum_{y=r+1}^{r+Q} G^T(y, K)}{Q} \quad (20)$$

where

$$G^T(y, K) = h(y - \mu) + (h + \hat{\pi}_1)b_1(y, K) + (h + \hat{\pi}_2)b_2(y, K) + \lambda_1\pi_1a_1(y, K) + \lambda_2\pi_2a_2(y, K) \quad (21)$$

Chen and Zheng (1993) provide conditions for the convexity of the loss function $G(\cdot)$ in a traditional (Q, r) model. The following theorem extends their result to a (Q, r, K) policy operating under threshold clearing.

Theorem 1 $G^T(y, K)$ is convex in y if $\lambda\pi_1 \leq (h + \hat{\pi}_1)$ and $\lambda\pi_2 \leq (h + \hat{\pi}_2)$.

The proof of this theorem (given in the Appendix) follows naturally from the structural results presented in section 4.3.

The form of $C^T(Q, r, K)$, given by equation (20), allows us to use a variant of the efficient algorithm proposed by Federgruen and Zheng (1992) to find the optimal policy parameters (Q^T, r^T, K^T) . Before describing this algorithm, we offer the following intuitive result

Theorem 2 If $\hat{\pi}_1 = \hat{\pi}_2$ and $\pi_1 = \pi_2 = 0$, then the optimal threshold rationing level $K^* = 0$.

Theorem 2 implies that if both class 1 and class 2 penalty costs are equal then there is no benefit to rationing and our policy simplifies to a standard (Q, r) system. This is not surprising, since the role of rationing is to provide priority to class 1 customers based on their assumed higher service needs.

We are now ready to define our solution approach. Note that for a fixed Q and K , the cost function consists of the sum of Q values of the function $G(\cdot)$. The unimodality of $-G(\cdot)$ implies that for fixed Q and K , $C^{T*}(Q, K) = \min_r C^T(Q, r, K)$ is achieved when the sum in equation (20) consists of the Q smallest values of this function; and these values are achieved in Q contiguous points. Also, $Q^{T*}(K)$, the optimal order size for a given K , is the largest value of Q for which $C^{T*}(Q - 1, K) > G_{Q,K}^T$ with $G_{Q,K}^T$ being the Q^{th} smallest $G^T(\cdot, K)$ value.

Let $y_{q,K}$ be the q^{th} smallest value of the function $G^T(y, K)$. Also, let $L(q, K) = \min\{y_{1,K}, y_{2,K}, \dots, y_{q,K}\}$ be the smallest of these q values, and $R(q, K) = \max\{y_{1,K}, y_{2,K}, \dots, y_{q,K}\}$ be the largest of these q values. The following two Lemmas extend properties derived by Federgruen and Zheng (1992) for a traditional (Q, r) policy, to our (Q, r, K) framework.

Lemma 6 For any $Q \geq 1, K \geq 0, r^*(Q, K) = L(Q, K) - 1$.

Lemma 7 $Q^*(K)$ is the smallest integer q with the property $C^*(q, K) \leq G(y_{q+1,K})$.

Using Lemma 7, the optimal reorder levels $r^*(1, K), r^*(2, K), \dots, r^*(Q, K)$ for given order quantities $1, \dots, Q$ are identified by the following procedure.

1. Let $q = 1, y_{1,K} = \min_y G(y, K)$, and

$$L(1, K) = R(1, K) = y_{1,K}$$

2. Increment $q = q + 1$, and compute

$$y_{q,K} = \begin{cases} L(q-1, K) - 1 & \text{if } G(L(q-1, K) - 1) \leq G(R(q-1, K) + 1) \\ R(q-1, K) + 1 & \text{otherwise} \end{cases}$$

Also set

$$L(q, K) = \min\{y_{1,K}, y_{2,K}, \dots, y_{q,K}\}$$

$$R(q, K) = \max\{y_{1,K}, y_{2,K}, \dots, y_{q,K}\}$$

3. if $q < Q$ GOTO step 2.

$$\text{else } r^*(Q, K) = L(Q, K) - 1$$

STOP

Using the search method above the optimal $r^T(Q, K)$, for given values of Q and K , is easily found. The optimal value of Q for a given K is then found by incrementing Q until the condition in Lemma 8 is satisfied. Federgruen and Zheng (1992) provides further details on this sequential procedure. Finally, the optimal K is found by performing a complete search on all possible values.

5 Comparing Clearing Mechanisms

Recall that the model in section 4 assumes a threshold clearing mechanism is used to clear the backlog when a replenishment order arrives. To test how well this model approximates the more attractive “priority clearing” mechanism (where class 1 backorders are always cleared before class 2 backorders), we conducted a numerical study. A program was written in C to simulate the priority clearing mechanism, as an analytical analysis is not possible. A sufficiently large sample of demand arrivals (approximately 10,000) was used in each case to ensure stability of the estimates. The simulation was run for a wide range of (Q, r, K) parameters and the parameters which gave the least cost were identified as optimal.

Let (Q^j, r^j, K^j) denote the optimal policy parameters under mechanism j , where $j = P$ (priority clearing), or T (threshold clearing). Note that for $j = T$ the parameters are truly optimal (as derived in section 4), while for $j = P$ the parameters are the result of an exhaustive search. To compare these solutions, we generated 54 problem sets varying in setup cost, ratio of class 1 versus total demand, and ratio of class 2 versus class 1 penalty cost. Our order setup cost scenarios were chosen, based on our interaction with various industries (Cohen et al., 1997, 1998, 2002), to reflect three different industry categories: high tech industries, such as Aerospace, Defense

$\hat{\pi}_2/\hat{\pi}_1$	Setup Cost = \$200		Setup Cost = \$100		Setup Cost = \$ 0	
	% Gap	% Gap	% Gap	% Gap	% Gap	% Gap
	Threshold	Hybrid	Threshold	Hybrid	Threshold	Hybrid
	vs	vs	vs	vs	vs	vs
	Priority	Priority	Priority	Priority	Priority	Priority
0.05	0.00%	0.00%	0.00%	0.00%	5.98%	3.00%
0.10	0.00%	0.00%	0.00%	0.00%	5.10%	0.00%
0.20	0.00%	0.00%	0.00%	0.00%	3.30%	0.00%
0.25	0.00%	0.00%	0.00%	0.00%	2.50%	0.00%
0.50	0.00%	0.00%	0.00%	0.00%	1.20%	0.00%
1.00	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 1: Comparison of Threshold(C^T), Hybrid and Priority Clearing (C) Mechanisms for $\lambda_1 = \lambda_2$

and Semi-conductor equipment, having extremely high setup costs ($s = \$200$), computers and telecom industries with more moderate setup costs ($s = \$100$), and commodity and packaged goods industries who enjoy little or no setup cost ($s = \$0$). Setup cost has a significant impact on most ordering schemes, particularly in the choice of order quantity (e.g., $s = \$0$ implies an optimal order quantity of one). Within each of these setup cost categories, demand and penalty cost scenarios were chosen to capture a broad range of customer environments. The demand ratios reflect cases where class 1 demand is less than, equal to, and greater than class 2 demand ($\lambda_1/\lambda = 0.25, 0.5, 0.75$). The ratio of class 2 to class 1 penalty costs range from severe cost differences to no difference ($\hat{\pi}_2/\hat{\pi}_1 = 0.05, 0.1, 0.2, 0.25, 0.5, 1$, with $\hat{\pi}_1$ held fixed at \$6000). We chose a wide range of stockout ratios since this difference triggers the degree of rationing needed in the inventory policy. The lead-time τ was assumed to be 3 months throughout with holding costs of $h = \$250$, which is typical of the military environment we studied (Cohen et al. 1998).

Table 1 compares the cost under the threshold and priority clearing mechanisms for the equal demand case (i.e., $\lambda_1/\lambda = 0.5$). The first column shows the percentage cost gap between the threshold clearing $C^T(Q^T, r^T, K^T)$ and priority clearing $C(Q^P, r^P, K^P)$ mechanisms. The second column lists the cost gap between the priority clearing mechanism using the least cost threshold clearing parameters (Q^T, r^T, K^T) (henceforth referred as “hybrid” mechanism), and the least cost priority clearing mechanism $C(Q^P, r^P, K^P)$. We found very little difference between the least cost threshold clearing policy and the least cost priority clearing policy when setup costs are medium to

high. In these cases the optimal Q is large compared to mean-lead time demand, so the probability of carrying over class 2 backorders from one replenishment period to the next is small. This limits the possible error introduced by the threshold clearing scheme. In contrast, when the setup cost is zero, the optimal Q is much smaller than mean lead-time demand. In this case, there is a performance gap between the threshold clearing and the priority clearing mechanisms, which decreases as the two penalty costs diverge. Interestingly, in all but one scenario, the optimal parameters are identical for both policies (i.e., $Q^T = Q^P, r^T = r^P, K^T = K^P$). Even for the one scenario where the parameters differed, the additional cost of using parameters (Q^P, r^P, K^P) , rather than (Q^T, r^T, K^T) , within a priority clearing scheme is only 3%.

Deshpande (2000) provides tables, similar to Table 1, for the other demand ratios of 0.25 and 0.75. The maximum gap between the least cost threshold clearing policy and the least cost priority clearing policy in these cases was 6.8%. The maximum gap between the cost of the least cost priority clearing policy and the cost of the priority clearing policy using (Q^T, r^T, K^T) was 3.3%. This occurred for parameters $\lambda_1/\lambda = 0.25$, $s = \$0$, and $\hat{\pi}_2/\hat{\pi}_1 = 0.05$.

In practice, we recommend using the policy parameters (Q^T, r^T, K^T) determined by our model, but clearing backlogs according to the priority clearing mechanism (“hybrid” mechanism). Our results suggest that this closely mimics the optimal (Q, r, K) policy under priority clearing in most cases. We observe a small gap in performance only when order setup costs are extremely small (typical of commodity industries) and penalty costs for the two demand classes are significantly different. In the next section, we compare this “hybrid” policy to several other policies used in practice.

6 Comparing Policy Performance

In this section we compare the cost of our hybrid (Q, r, K) policy to traditional round-up and separate stock policies, as well as to a lower bound over all possible rationing policies.

6.1 (Q, r, K) versus Traditional Static Policies

To test the cost effectiveness of the rationing policy, we used the same 54 problem sets outlined in section 5. As expected, our hybrid policy outperformed both the round-up and separate stock policies in all cases. A more interesting question is under what conditions does the (Q, r, K) policy provide the most benefit relative to these traditional policies.

$\hat{\pi}_2/\hat{\pi}_1$	% Benefit vs Round-up			% Benefit vs Separate Stock		
	$s = \$200$	$s = \$100$	$s = \$0$	$s = \$200$	$s = \$100$	$s = \$0$
0.05	17.78%	23.03%	37.68%	34.15%	34.08%	37.72%
0.1	13.28%	16.62%	30.25%	34.55%	34.17%	42.10%
0.2	8.48%	12.13%	18.60%	36.70%	38.78%	42.28%
0.25	7.38%	10.20%	15.16%	37.92%	39.26%	42.50%
0.5	4.51%	5.72%	8.94%	42.86%	42.88%	47.06%
1	0.00%	0.00%	0.00%	43.76%	44.46%	47.86%

Table 2: Benefit of Hybrid (Q, r, K) Policy vs Round-up and Separate Stock Policies

Table 2 provides some insight into this question by reporting the percent benefit (i.e., percent decrease in expected cost) of the hybrid (Q, r, K) policy versus the round-up and separate stock policies for the equal demand case ($\lambda_1/\lambda = 0.5$). Overall, its benefit is greater relative to the separate stock policy, with cost reductions of 34.15% to 47.86%. Its benefit over the round-up policy is more sensitive to the values of s and $\hat{\pi}_2/\hat{\pi}_1$, with reductions ranging from 0% to 37.68%.

The main advantage of a (Q, r, K) policy over round-up is its ability to provide differentiated service to the lower cost, class 2 customer. Consequently, we would expect a (Q, r, K) policy to offer the most benefit when the class 2 delay cost is significantly less than the class 1 delay cost. In Table 2, we see that the benefit of our hybrid policy over a round-up policy does indeed increase as the two delay costs diverge (i.e., as $\hat{\pi}_2/\hat{\pi}_1$ decreases). In contrast, the main advantage of a (Q, r, K) policy over a separate stock policy is its ability to pool inventory and thus offer the same differentiated service with less inventory investment. These pooling benefits are most pronounced when the two delay costs are the same (i.e., $\hat{\pi}_2/\hat{\pi}_1 = 1$) since no rationing occurs in this case (see Theorem 2). Table 2 confirms that the benefit of our hybrid policy over a separate stock policy is indeed greatest when $\hat{\pi}_2/\hat{\pi}_1 = 1$. As the delay costs diverge (i.e., $\hat{\pi}_2/\hat{\pi}_1$ decreases), the hybrid (Q, r, K) policy chooses to pool less inventory (i.e., increase its threshold level K) and thus its benefit over the separate stock policy, while still significant, decreases.

It is interesting to note that the benefit of a (Q, r, K) policy over either traditional policy appears to increase as the order setup cost decreases. This is because when setup costs are high, batch size increases and cycle length increases. As a result a high level of service is provided to

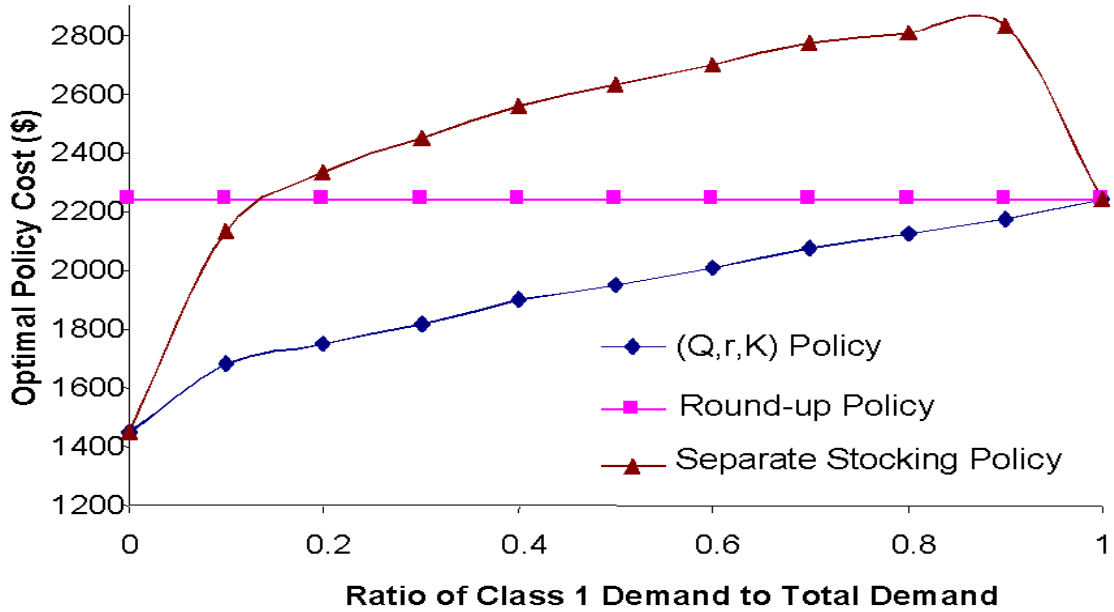


Figure 4: Optimal Policy Cost vs $\lambda_1/(\lambda_1 + \lambda_2)$

all customer classes and the effect of the reorder levels on cost will be relatively small. In this environment, the effect of service differentiation is less important. Conversely, when setup costs are low, batch sizes are low and the reorder level has a greater impact on overall cost. In this case the round-up policy causes a significant increase in reorder levels, leading to higher inefficiencies. Similarly, when setup costs are low, using two separate higher reorder levels causes the separate stock policy to be inefficient compared to a (Q, r, K) policy.

The general trends illustrated in Table 2 also hold for the other values of (λ_1/λ) in our data set. Figure 4 provides some additional insight into how the benefit of our proposed policy varies with (λ_1/λ) by graphing the costs of the three policies for different demand ratios, changing the percentage of demand attributed to class 1 customers while keeping total demand constant. The separate stock policy, intuitively, is identical to the hybrid (Q, r, K) policy when demand consists entirely of one customer type (i.e., $\lambda_1/\lambda = 0$ or 1). Figure 4 shows that the hybrid policy is beneficial, relative to the separate stock policy, in all but these two extreme cases. In fact, the cost of the separate stock policy increases significantly, relative to the hybrid policy, once even a small percentage of class 1 or 2 customers enter the mix. In contrast, the round-up policy is identical to the hybrid (Q, r, K) policy only when demand consists entirely of class 1 customers (i.e., $\lambda_1/\lambda = 1$). As the percentage of class 1 customers decreases, the benefit of using a (Q, r, K) policy over a round-up policy increases monotonically. This is because the hybrid (Q, r, K) policy

saves on inventory as the percentage of class 2 customers increase and more customers tolerate a lower service level. The round-up policy is most inefficient when class 1 demand is relatively small, since here it supports a large fraction of demand at a higher service level than needed.

These numeric results suggest that our hybrid (Q, r, K) policy offers a significant cost benefit over traditional round-up policies when delay costs are significantly different between classes, and demand consists of a large proportion of class 2 customers. These conditions are quite broad and apply to a large number of industries including commodity, and some segments of high tech. The policy offers little benefit when both the proportion of class 1 demand is high and the delay costs are roughly the same for the two classes. Turning to the comparison with traditional separate stock policies, our proposed policy offers a significant benefit as long as there is a reasonable mix of customer classes (i.e., $\lambda_1/\lambda \neq 0$ or 1). This benefit is substantial in most cases and increases in magnitude with $\hat{\pi}_2/\hat{\pi}_1$ and decreasing setup cost.

Before closing our discussion of traditional policies, it is worth pointing out that the benefits of pooling and differentiation offered by a (Q, r, K) policy will change if demand for the two customer classes are correlated. For example, Table 3 reports the benefit of the hybrid (Q, r, K) policy when demand is perfectly correlated. Comparing Tables 3 and 2, we see that the potential benefit of a (Q, r, K) policy over a round-up policy is even greater when demand is perfectly correlated. The gains versus a separate stock policy are still significant, although possibly less than with independent demand, particularly when the two delay costs are similar (i.e., $\hat{\pi}_2/\hat{\pi}_1$ approaches 1).

$\hat{\pi}_2/\hat{\pi}_1$	% Benefit vs Round-up			% Benefit vs Separate Stock		
	$s = \$200$	$s = \$100$	$s = \$0$	$s = \$200$	$s = \$100$	$s = \$0$
0.05	38.72%	45.82%	62.76%	44.16%	34.94%	10.91%
0.1	27.58%	32.28%	45.35%	37.43%	29.07%	8.06%
0.2	18.10%	19.77%	28.50%	35.38%	25.79%	5.05%
0.25	14.35%	16.92%	23.31%	33.66%	25.39%	3.98%
0.5	6.90%	7.44%	11.09%	32.94%	23.23%	2.18%
1	0.00%	0.00%	0.00%	30.77%	22.59%	0.76%

Table 3: Benefit of Hybrid (Q, r, K) Policy vs Round-up and Separate Stock Policies for Perfectly Correlated Demand

These results were developed using a correlated Poisson processes modeled by the arrival vector

$(\lambda_1, \lambda_2, \lambda_{12})$ (Xu, 1999). Here λ_1 and λ_2 represent the independent arrival rates for the two customer classes, and λ_{12} represents the arrival rate for the process where a demand arrival consists of both class 1 and class 2 demand. Note that in this formulation the class 1 and class 2 arrivals are still Poisson processes, although no longer independent because of the simultaneous arrival rate λ_{12} . Our analysis in earlier sections assumed that $\lambda_{12} = 0$. For perfectly correlated Poisson processes, the independent arrival rates for the two customer classes are zero and λ_{12} indicates the rate of joint arrivals of the two customer classes (i.e. $\lambda_1 = 0, \lambda_2 = 0, \lambda_{12}$). A C program was written to simulate the perfectly correlated arrival process. The optimal (Q, r, K) for the correlated process was identified by performing an exhaustive search.

Since performing such an exhaustive search may be computationally prohibitive for larger problems, it is interesting to see how much error one would introduce by using the model developed in section 4 to set the parameters Q, r , and K . Table 4 shows the % cost gap between the cost of our proposed “hybrid” threshold rationing policy assuming independent demand arrivals and the optimal cost for the perfectly correlated demand process with $\lambda_{12} = 10$. The cost gap decreases with both the order setup cost and the class 2 penalty cost. The gap is very small when both setup cost and $\hat{\pi}_2/\hat{\pi}_1$ are small, while the gap is as high as 15% when both setup costs and $\hat{\pi}_2/\hat{\pi}_1$ are large. An interesting observation of our numerical study is that our the hybrid model chooses the correct threshold rationing level K for all problem sets, but overestimates Q and underestimates r when demand is perfectly correlated. It appears that the (Q, r) model itself may lead to estimation errors for perfectly correlated processes, while the presence of rationing dampens that effect for low values of setup cost and class 2 penalty costs.

	% Cost Gap	% Cost Gap	% Cost Gap
$\hat{\pi}_2/\hat{\pi}_1$	Setup Cost = \$200	Setup Cost = \$100	Setup Cost = \$ 0
0.05	1.81%	1.77%	0.00%
0.1	1.41%	2.16%	0.00%
0.2	2.32%	2.17%	0.00%
0.25	3.45%	3.82%	0.22%
0.5	5.16%	5.03%	0.50%
1	15.12%	8.56%	3.07%

Table 4: Comparison of Hybrid versus Optimal (Q, r, K) for Perfectly Correlated Poisson Process

6.2 (Q, r, K) versus a Lower Bound on Optimal Rationing Policies

To gain insight into how a (Q, r, K) policy performs relative to other possible policies, we derive a lower bound on the optimal cost of a rationing policy (in essence, a lower bound over all possible policies). Let $G^R(y)$ denote the loss function which indicates the rate at which inventory holding and backorder costs accumulate at time $t + \tau$ given that the inventory position was y at time t under a rationing policy R in steady state. Recall that for a (Q, r) ordering policy the limiting inventory position is uniformly distributed between $r + 1$ and $r + Q$. Hence, the cost function for any policy R can be written as follows:

$$C^R(Q, r) = \frac{s\lambda + \sum_{y=r+1}^{r+Q} G^R(y)}{Q} \quad (22)$$

We seek to establish a lower bound on $G^R(y)$ by assuming perfect information over a lead-time. Suppose we knew the number of class 1 arrivals over the lead time. Then we would first fill these class 1 demands and then use the remaining inventory for class 2 demand. Thus a lower bound on $G^R(y)$ is obtained by assuming that all of y is reserved for class 1 demand and class 2 demand is filled after satisfying class 1 demand. We denote this lower bound by $G^l(y)$. For this lower bound we first compute the class 1 and 2 stockout probability, and class 1 and 2 backorder rates as follows:

The class 2 stockout probability under perfect information is given by

$$A_2^l(Q, r) = \frac{1}{Q} \sum_{y=r+1}^{r+Q} a_2^l(y) \quad (23)$$

where

$$a_2^l(y) = \sum_{x=y}^{\infty} p(x; \lambda\tau)$$

The long-run average number of class 2 backorders under perfect information is given by

$$B_2^l(Q, r) = \frac{1}{Q} \sum_{y=r+1}^{r+Q} b_2^l(y) \quad (24)$$

where

$$b_2^l(y) = \sum_{x=y}^{\infty} p(x; \lambda\tau) \left\{ \sum_{j=0}^{x-y} j b(\alpha_2; x; j) + \sum_{j=x-y+1}^x (x-y) b(\alpha_2; x; j) \right\}$$

Similarly the long-run fraction of time the system is out of stock for class 1 demand under perfect information is calculated by

$$A_1^l(Q, r) = \frac{1}{Q} \sum_{y=r+1}^{r+Q} a_1^l(y) \quad (25)$$

where

$$a_1^l(y) = \sum_{x=y}^{\infty} p(x; \lambda_1 \tau)$$

Also, the average number of class 1 backorders under perfect information is given by

$$B_1^l(Q, r) = \frac{1}{Q} \sum_{y=r+1}^{r+Q} b_1^l(y) \quad (26)$$

where

$$b_1^l(y) = \sum_{x=y}^{\infty} (x - y) p(x; \lambda_1 \tau)$$

Thus $G^l(y)$ can now be written as

$$G^l(y) = h(y - \mu) + (h + \hat{\pi}_1) b_1^l(y) + (h + \hat{\pi}_2) b_2^l(y) + \lambda_1 \pi_1 a_1^l(y) + \lambda_2 \pi_2 a_2^l(y) \quad (27)$$

We obtain a lower bound C^l by minimizing as follows:

$$C^l = \min_{Q, r} \frac{s\lambda + \sum_{y=r+1}^{r+Q} G^l(y)}{Q} \quad (28)$$

A C program was written to evaluate the above lower bound numerically. We identified the lower bound for the 54 problem sets outlined in section 5 by performing an exhaustive search. Table 5 shows the % cost gap between our hybrid (Q, r, K) policy and the lower bound for the equal demand case (i.e., $\lambda_1/\lambda = 0.5$, see section 5). The gap appears to increase as the setup cost decreases and the two penalty costs diverge. Recall that when the setup cost is small, the optimal order quantity is likely small compared to the mean lead-time demand. In this case, the threshold clearing mechanism is likely to clear some class 2 backorders before class 1 backorders. Also for very small class 2 penalty costs, the impact of clearing class 2 backorders before class 1 backorders is large for a threshold rationing policy. Hence the gap with the lower bound increases as the two penalty costs diverge. The worst case scenario (i.e., $s = 0, \hat{\pi}_2/\hat{\pi}_1 = 0.05$, which yielded a gap of 30.2%) is representative of firms in a commodity industry which offer a wide range of customer service options. Deshpande (2000) provides similar tables for the two other demand ratios (i.e., $\lambda_1/\lambda = 0.25, 0.75$). The largest gap among these problem sets was 34%. In our military study we found that the class 1 fillrates were around 95%, while the class 2 fillrates were around 85%, implying backorder costs of $\hat{\pi}_1 = \$6000$ and $\hat{\pi}_2 = \$1200$. A gap of 13% was observed for these parameters.

Note that our lower bound is derived by assuming perfect information over a lead-time which leads to the proper utilization of inventory between class 1 and class 2 customers. We doubt that

any feasible priority policy could achieve the costs derived by this “perfect information” lower bound. We suspect that the gap between our threshold rationing policy and the *actual* unknown non-stationary optimal policy is much less than the numbers in Table 5 suggest.

$\hat{\pi}_2/\hat{\pi}_1$	% Gap		
	Setup Cost = \$200	Setup Cost = \$100	Setup Cost = \$ 0
0.05	18.26%	23.00%	30.2%
0.1	17.37%	21.20%	27.1%
0.2	13.07%	15.30%	20.6%
0.25	11.87%	13.70%	17.00%
0.5	4.93%	6.27%	8.30%
1	0.00%	0.00%	0.00%

Table 5: Comparison of the Hybrid (Q, r, K) Policy vs a Lower Bound

7 Problem formulation for more than two demand classes

Although we limit our analysis in this paper to two demand classes, we envision that a (Q, r, K) type policy could be extended to handle any arbitrary number of classes. For example, suppose there are $i = 1, \dots, n$ demand classes ordered so that $i = 1$ denotes the highest priority and $i = n$ denotes the lowest priority. One can then envision a policy of the form $(Q, r, K_1, K_2, \dots, K_n)$ with nested rationing levels such that $K_n \geq K_{n-1} \geq \dots \geq K_2 \geq K_1 = 0$. In this nested rationing scheme, all demands of class $i \leq j$ are filled FCFS until the on-hand inventory level hits K_j (while class $i > j$ demands are backlogged). Once on-hand inventory hits K_j , newly arriving class j demands are backlogged as well. This nested rationing policy is similar to the multi-level rationing policies for make-to-stock production systems described by Ha (1997b) and Vericourt et al. (2002).

For $n > 2$, our objective function expands to

$$C^T(Q, r, K_1, \dots, K_n) = \frac{s\lambda + \sum_{y=r+1}^{r+Q} G^T(y, K_1, \dots, K_n)}{Q} \quad (29)$$

where

$$G^T(y, K_1, \dots, K_n) = h(y - \mu) + \sum_{i=1}^n (h + \hat{\pi}_i) b_i(y, K_1, \dots, K_n) + \sum_{i=1}^n \lambda_i \pi_i a_i(y, K_1, \dots, K_n). \quad (30)$$

The steady state probabilities a_i and b_i , for class $i = 1, \dots, n$, are much more difficult to calculate when $n > 2$. This is because a_i and b_i are now based on the *sequence* of demand arrivals of *all* n

classes. We leave a full study of this $(Q, r, K_1, K_2, \dots, K_N)$ policy for future research. However, we conjecture that the structural results shown in section 4.3 will hold when $n > 2$. For example, it is intuitive that increasing the reorder level r will decrease the backorder probabilities of all classes, while increasing the threshold level K_i will increase the backorder probabilities for classes $j \geq i$ and decrease the backorder probabilities of classes $j < i$. This is because increasing K_i reserves more inventory for demand classes with priority greater than class i .

8 Conclusions and Extensions

Motivated by a study of military logistics, we considered an inventory replenishment policy supporting two demand classes, differing in delay and shortage penalty costs and demand arrival rates. More specifically, we developed a model for selecting policy parameters and analyzing performance of a threshold rationing policy under a continuous review (Q, r) inventory framework. Our model includes some key practical features such as positive setup costs, positive lead-times and customer backorders in a continuous time framework, which have not been previously addressed in the literature.

By considering a more complete definition of backorder clearing, we derived closed form expressions for performance measures, such as average backorders and fillrate, for the given threshold rationing policy. Structural results on the sensitivity of performance measures to control parameters, as well as convexity results on the cost function, were also established. These results were used to formulate an efficient algorithm for computing the optimal policy parameters (Q, r, K) for the threshold rationing policy.

In addition to these analytical contributions, the paper offers the following important managerial insights:

1. *An estimation of the potential savings in switching from commonly used policies for differentiated supply chains to a threshold rationing policy.* Our analysis shows that a threshold rationing policy can significantly reduce inventory costs over current practice and at the same time provide the differentiated service required by customers. Our analysis provides a cost justification to upper level managers for moving to a new allocation policy.
2. *An understanding of the environments where the policy is most attractive.* In our numerical analysis we identify policy parameter ranges where a rationing policy could lead to significant cost savings. We show that the rationing policy offers significant savings over a separate stock policy as long as there is a reasonable mix of class 1 and 2 customers. The rationing

policy also offers significant savings over a round-up policy when delay costs are significantly different between classes, and demand consists of a large proportion of class 2 customers. Finally, the policy continues to offer savings when applied to environments with perfectly correlated demand. Managers can use the insights from this analysis to decide which products or parts should be managed by a rationing policy. The analysis also quantifies the expected cost savings over current policies for an item with given demand characteristics.

3. *A conservative bound on how well our simple rationing policy performs relative to an unknown, non-stationary “optimal” policy.* Our numerical results suggest that, for parameter values similar to the ones we observed in the military, the gap between our proposed policy and the unknown optimal non-stationary policy is less than 13%. The gap may be more significant for commodity industries (i.e., industries with low setup costs) which offer a wide range of customer service options (i.e., support a low $\hat{\pi}_2/\hat{\pi}_1$ ratio). Compared with potential non-stationary policies, our proposed policy also has the benefit of being relatively easy for managers to understand and implement.

Our two demand class model could also be extended in several other ways. More research needs to be carried out to establish analytical results for demand processes which are non-stationary, correlated or non-Poisson. A tighter lower bound on the optimal policy would also be helpful to provide more accurate comparisons. In addition, further analysis could be carried out for other forms of rationing policies. For example, a rationing policy where the rationing levels are non-stationary could be analyzed. Under such a policy, the rationing levels are a function of the on-hand stock and the arrival time of the next order. This is similar in spirit to Topkis’ periodic review model, where the rationing levels are revised every period. Such a policy would provide a finer level of service differentiation than the threshold rationing policy analyzed here.

Our rationing procedure could also be used to manage raw materials for products having both common and product specific components. A rationing policy in this case would allocate the common component between the two end product demands. The decision variables would be the ordering policies for the common and product specific components, and the rationing policy for the common component. In this case the end-product service level is determined by both the common and product specific component availability. Finally, we are currently working to extend our model to a decentralized environment where each player (supplier and customers) optimizes its individual objective function. This environment mirrors the decision making structure of DLA and the military services, and highlights the incentive problems they are still working to overcome.

References

- Alstrup, J., S. Boas, B. G. Madsen, and R. V. Vidal, "Booking Policy for Flights with Two Types of Passengers", *European Journal of Operations Research*, 1986, (27), 274-288.
- Baker, K., M. J. Magazine, and H. L. Nuttle, "The Effect of Component Commonality on Safety Stock in a Simple Inventory Model", *Management Science*, 1986, 32, 982-988.
- Belobaba, P., "Application of a Probabilistic Decision Model to Airline Seat Inventory Control", *Operations Research*, 1989, 37(2), 183-197.
- Bitran, G. R., and S. V. Mondschein, "An Application of Yield Management to the Hotel Industry with Multiple Day Stays", *Operations Research*, 1995, 43, 427-443.
- Bitran, G. R., and S. M. Gilbert, "Managing Hotel Reservations with Uncertain Arrivals", *Operations Research*, 1996, 44(1), 35-49.
- Chen, F., and Zheng, Y. S., "Inventory Models with General Backorder Costs", *European Journal of Operations Research*, 1993, 65, 175-186.
- Cohen, M. A., P. R. Kleindorfer, and H. L. Lee., "Service Constrained (s, S) Inventory Systems with Priority Demand Classes and Lost Sales", *Management Science*, 1988, 34(4), 482-499.
- Cohen, M. A., Y-S. Zheng, and V. Agrawal, "Service Parts Logistics: A Benchmark Analysis", *Technical Report*, The Wharton School, 1997, Philadelphia.
- Cohen, M. A., K. Donohue, and V. Deshpande, "Supply Chain Coordination Study: US Navy / Defense Logistics Agency", Project Report, *Fishman-Davidson Center for Service and Operations Management*, The Wharton School, February 1998, Philadelphia.
- Cohen, M. A., Ho, T., Ren, J. and Terwiesch, C., "Measuring Imputed Costs in the Semiconductor Equipment Supply Chain," *Working Paper*, The Wharton School, 2002, Philadelphia.
- Deshpande, V., "Supply Chain Coordination with Service Differentiated Customer Classes", *unpublished dissertation*, Dept. of Operations and Information Management, The Wharton School, University of Pennsylvania, 2000.
- Federgruen, A., and Y. Zheng, "An Efficient Algorithm for Computing an Optimal (r, Q) Policy in Continuous Review Stochastic Inventory Systems", *Operations Research*, 1992, 40(4), 808-813.
- Frank, K. C., R. Q. Zhang, and I. Duenyas, "Optimal Policies for Inventory Systems with Priority Demand Classes", *Working Paper*, University of Michigan, 1999.
- Gerchak, Y., and M. Henig, "An Inventory Model with Component Commonality", *Operations Research Letters*, 1986, 5, 157-160.
- Gerchak, Y., M. Magazine, and B. Gamble, "Component Commonality with Service Level Con-

straints”, *Management Science*, 1988, 34(6), 753-760.

Ha, A. Y., “Inventory Rationing in a Make-to-Stock Production System with Several Demand Classes and Lost Sales”, *Management Science*, 1997a, 43(8), 1093-1103.

Ha, A. Y., “Stock Rationing Policy for a make-to-stock production system with two priority classes and backordering”, *Naval Research Logistics*, 1997b, 43, 458-72.

Ha, A. Y., “Stock Rationing in an $M|E_k|1$ Make-to-Stock Queue”, *Management Science*, 2000, 46(1), 77-87.

Hadley, G., and T. Whitin, “Analysis of Inventory Systems”, *Prentice Hall*, Englewood Cliffs, NJ, 1963.

Jackson, P. L., “Stock Allocation in a Two-Echelon Distribution System or ‘What to Do Until Your Ship Comes In’ ”, *Management Science*, 1988, 34, 880-895.

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

Kimes, S. E., “Yield Management: A Tool for Capacity-Constrained Service Firms”, *Journal of Operations Management*, 1989, 8, 348-363.

Kleijn, M. J., and Dekker, R., “An overview of inventory systems with several demand classes”, *Econometric Institute Report 9838/A*, 1998, Erasmus University, Rotterdam, Netherlands.

Lee, T. C., and M. Hersh, “A Model for Dynamic Airline Seat Inventory Control with Multiple Seat Bookings”, *Transportation Science*, 1993, 27, 252-265.

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

Ross, K. W., and D. H. K. Tsang, “The Stochastic Knapsack Problem”, *IEEE Transactions on Communications*, 1989, 37, 740-747.

Savin, S., M. A. Cohen, N. Gans, Z. Katalan, “Capacity Management in Rental Businesses with Heterogeneous Customer Bases”, *Working paper*, 2000, The Wharton School, University of Pennsylvania.

Subramanian, J., S. Stidham, and C. Lautenbacher, “Airline Yield Management with Overbooking, Cancellations and No-shows”, *Transportation Science*, 1999, 33, 147-167.

Topkis, D. M., “Optimal Ordering and Rationing Policies in a Non-stationary Dynamic Inventory Model with n Demand Classes”, *Management Science*, 1968, 15(3), 160-76.

Veinott, A. F., “Optimal Policy in a Dynamic, Single Product, Non-stationary Inventory Model with Several Demand Classes”, *Operations Research*, 1965, 13, 761-778.

Vericourt, F. D., Karaesmen, F., and Dallery, Y., “Dynamic Scheduling in a Make-to-Stock Sys-

tem: A Partial Characterization of Optimal Policies”, *Operations Research*, 2000, 48(5), p811-819.

Vericourt, F. D., Karaesmen, F., and Dallery, Y., “Optimal Stock Allocation for a Capacitated Supply System”, 2002, *Working Paper*, Duke University, 2002.

Xu, S. H., “Structural Analysis of a Queueing System with Multiclasses of Correlated Arrivals and Blocking”, *Operations Research*, 1999, 47(2), p264-276.

Zhang, H., “A Note on the Convexity of Service-Level Measures of the (r, Q) System”, *Management Science*, 1998, 44(3), 431-432.

Zhao, W., and Y. S. Zheng, “A Dynamic Model for Airline Seat Allocation with Passenger Diversion and No-Shows”, *Transportation Science*, 2001, 35(1), p80-98.

Zheng, Y. S., “On Properties of Stochastic Inventory Systems”, *Management Science*, 1992, 38(1), 87-103.

Zipkin, P., “Stochastic Lead-times in Continuous Time Inventory Models”, *Naval Research Logistics Quarterly*, 1986a, 33, 763-774.

Zipkin, P., “Inventory Service-Level Measures: Convexity and Approximation”, *Management Science*, 1986b, 12(8), p975-981.

Appendix

Lemma 1 $A_2(Q, r, K)$ is decreasing in r and increasing in K .

Proof Using equation (13), it is sufficient to prove that $a_2(y, K)$ is decreasing in y and increasing in K . This is equivalent to showing that $\Delta_y a_2(y, K) \leq 0, \forall y$ and $\Delta_K a_2(y, K) \geq 0, \forall K$. Now for $y > K$

$$\Delta_K a_2(y, K) = a_2(y, K + 1) - a_2(y, K) = p(y - K - 1; \lambda\tau) \geq 0 \quad \forall y > K$$

Also, for $y \leq K, a_2(y, K + 1) - a_2(y, K) = 1 - 1 = 0$. Therefore $a_2(y, K)$ is increasing in K .

Now turning to the impact of y

$$\Delta_y a_2(y, K) = a_2(y + 1, K) - a_2(y, K) = -p(y - K; \lambda\tau) \leq 0 \quad \forall y \geq K$$

Also, for $y < K, a_2(y + 1, K) - a_2(y, K) = 1 - 1 = 0$. □

Lemma 2 $A_1(Q, r, K)$ is decreasing in K , and decreasing in r .

Proof Using equation (13), it is sufficient to prove that $a_1(y, K)$ is decreasing in y and K . This is equivalent to showing that $\Delta_y a_1(y, K) \leq 0, \forall y$ and $\Delta_K a_1(y, K) \leq 0, \forall K$. Now

$$\begin{aligned} \Delta_y a_1(y, K) &= a_1(y + 1, K) - a_1(y, K) \\ &= \sum_{x=y+1}^{\infty} p(x; \lambda\tau) \sum_{j=0}^{x-y-1} b(\alpha_1; x - y - 1 + K; K + j) - \sum_{x=y}^{\infty} p(x; \lambda\tau) \sum_{j=0}^{x-y} b(\alpha_1; x - y + K; K + j) \\ &= -p(y; \lambda\tau) b(\alpha_1; K; K) + \sum_{x=y+1}^{\infty} p(x; \lambda\tau) \left\{ \sum_{j=0}^{x-y-1} b(\alpha_1; x - y - 1 + K; K + j) - \sum_{j=0}^{x-y} b(\alpha_1; x - y + K; K + j) \right\} \\ &= -p(y; \lambda\tau) \alpha_1^K + \sum_{x=y+1}^{\infty} p(x; \lambda\tau) \left\{ \sum_{j=0}^{K-1} b(\alpha_1; x - y + K; j) - \sum_{j=0}^{K-1} b(\alpha_1; x - y - 1 + K; j) \right\} \\ &= - \sum_{x=y}^{\infty} p(x; \lambda\tau) \alpha_1 b(\alpha_1; x - y - 1 + K; K - 1) \\ &\leq 0 \end{aligned}$$

Therefore $a_1(y, K)$ is decreasing in y . Looking at the impact of K we can show that

$$\begin{aligned} \Delta_K a_1(y, K) &= a_1(y, K + 1) - a_1(y, K) \\ &= \sum_{x=y}^{\infty} p(x; \lambda\tau) \sum_{j=0}^{x-y} b(\alpha_1; x - y + K + 1; K + 1 + j) - \sum_{x=y}^{\infty} p(x; \lambda\tau) \sum_{j=0}^{x-y} b(\alpha_1; x - y + K; K + j) \\ &= \sum_{x=y}^{\infty} p(x; \lambda\tau) \left\{ \sum_{j=0}^{K-1} b(\alpha_1; x - y + K; j) - \sum_{j=0}^K b(\alpha_1; x - y + K + 1; j) \right\} \end{aligned}$$

$$\begin{aligned}
&= - \sum_{x=y}^{\infty} p(x; \lambda\tau)(1 - \alpha_1)b(\alpha_1; x - y + K; K) \\
&\leq 0
\end{aligned}$$

□

Lemma 3 $B_2(Q, r, K)$ is decreasing in r and increasing in K .

Proof Using equation (15), it is sufficient to prove that $b_2(y, K)$ is decreasing in y and increasing in K . Now for $y > K$

$$\Delta_K b_2(y, K) = b_2(y, K + 1) - b_2(y, K) = \alpha_2 P(y - K; \lambda\tau) \geq 0$$

Also for $y \leq K$, $\Delta_K b_2(y, K) = \alpha_2 \geq 0$. Similarly, for $y \geq K$

$$\Delta_y b_2(y, K) = b_2(y + 1, K) - b_2(y, K) = -\alpha_2 P(y - K + 1; \lambda\tau) \leq 0$$

Also, for $y < K$, $\Delta_K b_2(y, K) = -\alpha_2 \leq 0$.

□

Lemma 4 $B_1(Q, r, K)$ is decreasing in r and decreasing in K .

Proof Using equation (15), it is sufficient to prove that $b_1(y, K)$ is decreasing in y and decreasing in K . Now

$$\begin{aligned}
\Delta_K b_1(y, K) &= b_1(y, K + 1) - b_1(y, K) \\
&= \sum_{x=y}^{\infty} p(x; \lambda\tau) \sum_{i=1}^{x-y} \left\{ \sum_{j=0}^{K+i-1} b(\alpha_1; x - y + K; j) - \sum_{j=0}^{K+i} b(\alpha_1; x - y + K + 1; j) \right\} \\
&= - \sum_{x=y}^{\infty} p(x; \lambda\tau) \sum_{i=1}^{x-y} (1 - \alpha_1) b(\alpha_1; x - y + K; K + i) \\
&\leq 0
\end{aligned}$$

Similarly one can show that,

$$\Delta_y b_1(y, K) = b_1(y + 1, K) - b_1(y, K) = - \sum_{x=y+1}^{\infty} p(x; \lambda\tau) \sum_{i=0}^{x-y-1} \alpha_1 b(\alpha_1; x - y - 1 + K; K + i) \leq 0$$

□

Lemma 5 $b_i(y, K)$ is convex in y for fixed K and convex in K for fixed y .

Proof These results follow directly from the derivations of Lemmas 1-4. □

Theorem 1 $G(y, K)$ is convex in its parameter y if $\lambda\pi_1 \leq (h + \hat{\pi}_1)$ and $\lambda\pi_2 \leq (h + \hat{\pi}_2)$.

Proof We show that the second difference for $G(\cdot, K)$ is positive under the stated conditions. Now

for $y > K$

$$\begin{aligned}
& \{G(y+1, K) - G(y, K)\} - \{G(y, K) - G(y-1, K)\} \\
&= (\alpha_2(h + \hat{\pi}_2) - \lambda_2\pi_2)p(y-K; \lambda\tau) + \lambda_2\pi_2p(y-K-1; \lambda\tau) \\
&+ \lambda_1\pi_1 \sum_{x=y-1}^{\infty} p(x; \lambda\tau)\alpha_1b(\alpha_1; x-y+K; K-1) \\
&+ (\alpha_1(h + \hat{\pi}_1) - \lambda_1\pi_1) \sum_{x=y}^{\infty} p(x; \lambda\tau)\alpha_1b(\alpha_1; x-y-1+K; K-1) \\
&\geq 0 \quad \text{if } \lambda\pi_1 \leq (h + \hat{\pi}_1) \quad \text{and} \quad \lambda\pi_2 \leq (h + \hat{\pi}_2)
\end{aligned}$$

Also for $y < K$

$$\begin{aligned}
& \{G(y+1, K) - G(y, K)\} - \{G(y, K) - G(y-1, K)\} \\
&= \lambda_1\pi_1 \sum_{x=y-1}^{\infty} p(x; \lambda\tau)\alpha_1b(\alpha_1; x-y+K; K-1) \\
&+ (\alpha_1(h + \hat{\pi}_1) - \lambda_1\pi_1) \sum_{x=y}^{\infty} p(x; \lambda\tau)\alpha_1b(\alpha_1; x-y-1+K; K-1) \\
&\geq 0 \quad \text{if } \lambda\pi_1 \leq (h + \hat{\pi}_1) \quad \square
\end{aligned}$$

Theorem 2 If $\hat{\pi}_1 = \hat{\pi}_2$ and $\pi_1 = \pi_2 = 0$, then the optimal threshold rationing level $K^* = 0$.

Proof Using Lemmas (1)-(4) it is easy to show that $b_1(y, K) + b_2(y, K)$ is increasing in K . Thus, for the symmetric cost structure, it is easy to see from equation (21) that $K^* = 0$. \square