

Spare Parts Inventory Management with Demand Lead Times and Rationing

Yasar Levent Kocaga
Alper Sen

Department of Industrial Engineering
Bilkent University
Bilkent, Ankara 06800, Turkey
{yasarl,alpersen}@bilkent.edu.tr

Revised April, 2005

Abstract

We study the spare parts service system of a major capital equipment manufacturer facing two kinds of orders of different criticality. The more critical down orders need to be supplied immediately, whereas the less critical maintenance orders allow a given demand lead time to be fulfilled. For this system, we propose a policy that rations the maintenance orders. Under a one-for-one replenishment policy with backordering and for Poisson demand arrivals for both classes, we first derive expressions for the service levels of both classes and then conduct a computational study to illustrate superior system performance compared to a system without rationing. We also conduct a case study with 64 representative parts and show that significant savings (as much as 14 % on inventory on hand) are possible through incorporation of demand lead times and rationing.

1 Introduction

The primary motivation behind this research is our experience with a leading capital equipment manufacturer. The company owns research, development, and manufacturing facilities in the United States, Europe, and the Far East and distributes its systems across the globe. The company is at the top of the supply chain for many high technology products.

The systems that the company is manufacturing are very expensive investments and are highly critical to the operations of its customers. Unused capacity at the customer manufacturing facilities due to equipment failures is very costly. In order to provide spare parts and service to customers for equipment failures and scheduled maintenance, the company has an extensive spare parts network. The network consists of more than 70 locations across the globe, that are the sites of company-owned distribution centers and depots. In addition, the company also has agreements with its leading customers to manage the stock rooms in customer facilities. Three regional distribution centers: one

in North America, one in Asia, and one in Europe constitute the backbone of the network and are primarily responsible for procuring and distributing spare parts to depots and customer locations. The depot locations are such that they can provide a 4-hour service to customers (those who do not have stock rooms operated by the company) for equipment failures (“down orders”). The regional distribution centers may also be used as a primary source for down orders for certain customers. In addition, the regional distribution centers provide a second level of support for down orders that cannot be satisfied from the local depots. Customers also demand spare parts to be used in their scheduled maintenance activities (“lead time orders”). The primary source to meet these demands is usually the regional distribution centers. However local depots can also be used for this purpose for certain customers. We have to state here that even though the maintenance activities are scheduled and thus are known in advance for the customers, the capital equipment manufacturer we study does not have visibility to such schedules. Since each location supports many different customers and a large install base, the capital equipment manufacturer perceives these orders as random.

Both types of customer orders (down and lead time) go through an order fulfillment engine which searches for available inventory in different locations according to a search sequence specific to each customer. Down orders need to be satisfied immediately (their request date is the date of order creation), whereas the lead time orders need to be satisfied at a future date. A depot may be facing down and lead time demand from a variety of customers, while a regional distribution center may be facing down and lead time demand from external customers in addition to the “replenishment orders” requested by internal customers: the depots and stock rooms managed by the company. The operation of this complex network is further complicated by a vast number of parts composed of consumables and non-consumables (more than 50,000 active parts need to be managed) and varying service level requirements by different customers.

While providing an implementable and “good” solution for the whole spares network is a proven challenge, we focus on an important issue where improvements can provide immediate and significant benefits. In the existing practice, for those locations that are facing different types of demand (down, lead time or replenishment), the company aims to achieve the maximum of the service level requirements while considering the aggregated demand. Moreover, the company does not recognize the possible demand lead times (the difference between requested date and ship date in excess of transportation time) for lead time orders and possible slacks (the difference between the replenishment lead time the company uses for planning downstream locations and transportation lead time) for replenishment orders. Obviously, this approach is inefficient. We suggest an inventory model that recognizes both demand lead times and multiple demand classes, and allows for providing differentiated service levels through rationing.

Multiple demand classes occur naturally in many inventory systems. Consider a two-echelon supply network consisting of a warehouse at the upstream and a number of retailers at the downstream. If the retailers are located in different regions and have different demand characteristics, it may be beneficial to assign retailers different priorities and differentiate demand accordingly. A similar example can be a two-echelon supply network where the upstream is a warehouse which supplies customers (directly) and the downstream retailers (in the form of replenishment orders). In such a case, the stock-out cost resulting from not being able to supply customers is usually much higher than that of the retailers, since the latter one causes only a delay in the replenishment orders, which usually results in a lower cost.

Multiple demand classes also occur in systems where an item can have different uses. In a production system, a part may be installed in various equipment that is crucial to the flow of production, while it may also be used in other equipment, downtimes of which might be less critical. Thus, the demand for this spare part can be differentiated into several demand classes. Again, in a production system where the same component is used in multiple end products of different value (based on measures such as profitability), the demand of the end products can be differentiated accordingly. Observe that, in both examples, the demand does not necessarily come from different end customers. Yet, multiple demand classes occur naturally in both examples either in the form of demand for a spare part from equipment of different criticality or demand for a common component from different end products.

Multiple demand classes can also be observed in other systems. Revenue management is a celebrated example. The underlying assumption here is that some customers are willing to pay more for a room or seat than are others. Therefore, it can be optimal to refuse a low-price customer in anticipation of a future request from a high-price customer. It is indeed optimal if the customers arrive sequentially (low-price customers followed by the high-price customers) and the optimal policy has shown to be characterized by a set of protection levels, which essentially are the inventory reserved for future (high-price) classes. Observe that, in these problems the inventory is perishable and this leads to non-stationary control policies which adjust as time to expiration (i.e., flight date of the plane) approaches. Another distinguishing fact is that inventory level (capacity) is fixed. Thus, as opposed to most classical inventory systems, the replenishment decisions are irrelevant.

Given a system with multiple demand classes, the easiest policy would be to use different stockpiles for each demand class. This way, it would be very easy to assign a different service level to each class. Also, the practical implementation of this policy would be relatively easy. But the drawback is that no advantage would be gained from the so-called portfolio effect. In other words, the advantage of demand pooling from different demand sources would no longer be materialized. Therefore, as a result

of the increasing demand variability, more safety stock would be needed to ensure a minimum required service level. On the other hand, one could simply use the same pool of inventory to satisfy demand from various customer classes without differentiating them. In this case, the highest required service level would determine the total inventory needed and thus the inventory cost. The drawback of this policy is that higher service levels are offered to the rest of the demand classes, leading to increased inventory costs.

Rationing, or the so-called critical level policy, essentially lies between these two extremes. Rationing has proved to be effective to handle different demand classes with different stock-out costs or service levels. Kleijn and Dekker [14] provide a comprehensive study illustrating various examples in which multiple demand classes arise together with a literature review on the applications of rationing in such environments. We will explain rationing assuming that there are two demand classes. In this setting, a certain part of the stock is reserved for high priority demand. This amount is called the critical level; once the inventory level drops to this level, demand from the lower priority demand class is no longer satisfied. If demand not satisfied immediately is backordered, how to handle the replenishment orders is another problem. Obviously, if there is a backorder for a high priority customer upon the arrival of a replenishment order, it is optimal to use this replenishment order to satisfy this backorder. In addition, if there is a backorder for a low priority customer upon the arrival of a replenishment order and the inventory level is at or above the critical level, one should use this replenishment order to satisfy this backorder. However, if there is a low priority backorder and the inventory level is below the critical level, one can either satisfy this backorder or leave items in inventory in anticipation of more critical orders. The latter option is referred to as the priority clearing mechanism and has been proven to be optimal under specific conditions. Under general conditions, however, whichever of these is optimal depends on the problem settings.

Except for very specific cases, a simple critical level policy with a static critical level will not be optimal. An optimal policy should take the remaining time until the next replenishment arrival into account. As the booking limits adjust to the remaining time until expiration in revenue management, the critical level in a rationing policy should also adjust dynamically. For example, if the inventory level is below the critical level, but it is known that a replenishment order will arrive within a short period of time, it may not be optimal to refuse a low priority demand arrival, especially if the probability of a high priority demand arrival within this time is very small. But employing such a dynamic rationing policy would be extremely difficult from a practical point of view. Thus, we prefer to focus on a static rationing policy where the critical level does not change over time.

The structure of the firm we study by itself exhibits different demand classes (down orders versus lead time orders) thereby creating an environment where rationing can be applied. Thus, our approach

in this paper incorporates rationing to the current practice of the firm with two demand classes differentiated by their demand lead-time. Our motivation in taking this approach is that we believe it will result in better system performance given certain service level requirements. We consider the down orders as the high priority (or critical) class and the maintenance orders as low priority (or non-critical).

While the specific application we study this problem requires a higher service level for the demand class that has no demand lead time, it is possible that other applications require lower service level for this demand class. Consider, for example, a multi-channel retailer which sells its goods online as well as through a bricks-and-mortar store. Online customers submit their orders in advance and a commitment is made upon the acceptance of these orders, whereas no prior commitment is made to the customers in the demand class without demand lead time, who ask for inventory upon their arrival to the store. Obviously, the service level requirement for online customers would be higher than customers purchasing through the store. Our general model allows rationing either type of demand class. We refer to the class that has higher (lower) service level requirement as critical (non-critical) class.

We will first model the system as a single location system facing a Poisson demand in two classes, class 1 and class 2 with rates λ_1 and λ_2 , respectively. The spare part inventory is replenished according to an $(S - 1, S)$ policy, S being the order-up-to level. For simplicity, we consider a deterministic replenishment lead-time, L . Class 2 demands have a deterministic demand lead-time of T while the class 1 orders must be satisfied immediately. The service level we consider will be the type I service level, i.e., the probability of no stock-out. Under these circumstances the policy works as follows: once a critical order comes, it is either satisfied (at its due date) or backlogged if there is no inventory. On the other hand, a non-critical order is satisfied only if the inventory level is above a critical level, S_c , otherwise it is backlogged. We assume that class 2 orders are always accepted and a delivery commitment is made for them at their due date. Our aim will be to find the optimum S and S_c such that the given service levels requirements $\bar{\beta}_1$ and $\bar{\beta}_2$ are satisfied.

The remainder of the paper is organized as follows. In Section 2, we provide a review of the literature of related inventory systems. In Section 3, we derive an exact expression for the non-critical customer class service level and an approximate expression for the critical customer class service level. We also show analytically that the approximate expression for the critical customer class service level is a lower bound for the actual service level. In addition, we present a service level optimization model and an algorithm to solve it. In Section 4, we present the results of our simulation study, which indicate that our approximation for the service level of the critical class works quite well for high service levels of the critical class. In addition, we present the results of the optimization study that we conducted

using our justified approximation for the critical service level. Also in Section 4, we present our results on a case study using 64 parts from the capital equipment manufacturer that we described earlier. In Section 5, we conclude the paper giving an overall summary of what we have done, our contribution to the existing literature and its practical implications.

2 Literature Review

In this section, we will first review the literature on inventory systems with demand lead time. Then we will elaborate on the literature about rationing. We find it useful to distinguish between the periodic review literature and continuous review literature. Therefore we will first focus on the periodic review models and then proceed with the continuous review models.

The concept of demand lead-time was first introduced by Simpson [19] by the term “service time” for base-stock, multi-stage production systems. Hariharan and Zipkin [12] then coined the name “demand lead-time” to describe inventory-distribution systems where customers do not require immediate delivery of orders and allow for a fixed delay. The key observation of both papers is that demand lead-time works just as the opposite of supply lead-time, reducing the inventory held for achieving the required service level. Obviously this fact also applies to the system we consider but the existence of the two service classes complicates the model. Moynadeh and Aggarwal [17] consider a two echelon system with two modes of inventory replenishment. However, in their model all orders are satisfied on a FCFS basis while the two order classes differ only in their transportation lead-times. On the other hand, we consider a system where orders are satisfied on a FDFS (first-due-first-serve) basis. Wang et al. [23] analyze a similar system in order to derive the transient and steady-state performance metrics of the system. This work is actually the most relevant to ours in terms of the presence of two classes of service differentiated by a demand lead-time. Therefore, we prefer to explore their work profoundly.

Wang et al. [23] first study a single location system and derive expressions for the inventory level distribution and random customer delay. As a result, an expected yet crucial observation is made: the service level of customers with positive demand lead times is higher than the service level for customers with zero demand lead time as long as there is a positive probability that the replenishment order corresponding to a customer with positive demand lead time arrives before its demand due date. After deriving the steady-state performance metrics for the single location system, the model is extended to a two echelon system. By following an approach similar to the well-known METRIC, the multi-echelon network is decomposed into single location subsystems. After the analysis of the two-echelon setting, it is seen that the system with two service classes results in significant inventory cost savings.

In the system we consider, the customers with positive demand lead times constitute the non-critical demand class, while the customers with zero demand lead times constitute the critical demand

class. Therefore, it is imperative that we use a policy that could provide a higher service level to the demand class with zero demand lead times. Rationing is such a policy. In the standard policy, whenever on-hand inventories drop below a certain level (usually called critical level, rationing level, or threshold level of the associated customer class) the demands of the lower priority classes are not satisfied with the expectation of future high priority class customer demands.

The literature about rationing begins with Veinott [22] who was the first to consider the problem of several demand classes in inventory systems. He analyzed a periodic review inventory model with n demand classes and zero lead-time with limited ordering, and introduced the notion of a critical level policy. Topkis [21] proved the optimality of this policy both for the cases of backordering and lost sales. The optimal rationing policy is such that demand from a given class is satisfied from existing stock as long as there remains no unsatisfied demand from a higher class and the stock level does not drop below a certain critical level for that class. The critical levels are generally decreasing with the remaining time until the next ordering opportunity. Independent of Topkis, Evans [7] and Kaplan [13] fundamentally derived the same results for two demand classes. Nahmias and Demmy [18] derived expressions for the expected backorder levels for a multi-period model with zero lead-times and an (s, S) inventory policy when a fixed critical level is used. Other work in periodic inventory models with multiple demand classes include Atkins and Katircioglu [1], Cohen et al. [2] and Frank et al. [8].

Nahmias and Demmy [18] were the first to consider multiple demand classes in a continuous review inventory model. They analyzed a (Q, r) inventory model, with two demand classes, Poisson demand, backordering, a fixed lead-time and a critical level policy, under the crucial assumption that there is at most one outstanding order. This assumption implies that whenever a replenishment order is triggered, the net inventory and the inventory position are identical. The model of Nahmias and Demmy is analyzed in a lost sales context by Melchioris et al. [16].

Deshpande et al. [6] considered a rationing policy for two demand classes differing in delay and shortage penalty costs with Poisson demand arrivals under a continuous review (Q, r) environment. They did not make the assumption of at most one outstanding order which makes the allocation of arriving orders a major issue to consider. They defined a so-called threshold clearing mechanism to overcome the difficulty of allocating arriving orders and provided an efficient algorithm for computing the optimal policy parameters which are defined by (Q, r, K) , K being the threshold level.

Dekker et al. [5] discussed a case study on the inventory control of slow moving spare parts in a large petrochemical plant, where parts were installed in equipments of different criticality. They studied a lot-for-lot inventory model with two demand classes, but without the assumption of at most one outstanding order. Demand for both classes is assumed to be Poisson while the replenishment lead-time is assumed to be deterministic. The primary contribution of this paper is the derivation of

service levels for both classes in the form of probability of no stock-out. However, the service level for the critical demand is only an approximation since it depends on how incoming replenishment orders are handled in a complicated way, while the service level for non-critical demand class is exact, since it is not effected by the way incoming orders are handled.

A relevant stream of research introduced by Ha [9] considers the limited production capacity for replenishment orders and analyzes the system through make-to-stock queues with multiple demand classes. Ha's [9] initial model has two demand classes, exponential supply lead times and backordering. Extensions include multiple demand classes (Vericourt et al. [3]), lost sales (Ha [10]), lost sales and Erlang lead time distributions (Ha [11]) and lost sales and general lead time distributions (Dekker et al. [4]).

Our study differs from earlier research in that we simultaneously consider demand lead times and rationing. We investigate a continuous time, single item, lot-for-lot model with backordering.

We finally note that most rationing papers including ours make the simplifying assumptions that there is a single-item in consideration and critical levels are time invariant. Notable extensions are two recent works. Kranenburg and van Houtum [15] considered a single location multi-item spare parts model with multiple demand classes. A solution procedure is developed based on Lagrange relaxation and 10-20 % savings on inventory investment are reported using real data from a semiconductor equipment manufacturer, ASML. Teunter and Haneveld [20] studied a critical level policy for two demand classes where the critical level depends on the remaining time until the next stock replenishment. The so-called remaining time policy is characterized by a set of critical stocking times $(L_1, L_2, \dots)$; if the remaining time until the next replenishment is at most L_1 , no items are reserved for the high-priority customers; if the time is between L_1 and $L_1 + L_2$ then one item should be reserved, and so on.

3 The Model

We consider a single location spare part inventory system which faces two classes of demand arrivals with different criticality. Class 1 and class 2 demand arrivals are both assumed to be Poisson with rates of λ_1 and λ_2 , respectively. Both arrivals are satisfied from the same pool of inventory which is controlled by a base stock policy with a base stock level S . Therefore, each demand arrival triggers a replenishment order with a deterministic lead time of L . In addition, the demand from type 2 class allows a deterministic demand lead time of T . Before proceeding with the description of our rationing policy, we provide the following notation which will be used throughout the rest of this paper:

λ_1	: Arrival rate of the demand class with no demand lead time (class 1);
λ_2	: Arrival rate of the demand class with demand lead time (class 2);
L	: Replenishment (supply) lead time;
T	: Demand lead time in class 2;
$\bar{\beta}_j$	: Service level requirement for class j ;
S	: Base stock level;
S_c	: Critical level;
$\beta_j^c(S, S_c)$	: Service level for the class j for given S and S_c if class j is the critical class;
$\beta_j^n(S, S_c)$	: Service level for the class j for given S and S_c if class j is the non-critical class;
$I(a)$	: Inventory level net of backorders for non-critical class at time a ;
$B(a)$	: Total backorders at time a ;
$D_j(a, b]$	: Class j demand due in interval $(a, b]$;
$R(a, b]$	: Replenishments that are received in interval $(a, b]$;
H	: Hitting time, i.e., the arrival time of the $(S - S_c)$ th total net demand.

Note that $D_j(a, b]$ is a Poisson random variable with rate $\lambda_j \times (b - a)$. In our model, we will use the type I service level, i.e., the probability of no stock-out, as our service level measure. We note that because of the PASTA (Poisson Arrivals See Time Averages) property, this is also the type II service level, i.e., the fill rate.

In our specific industrial application, we require $\bar{\beta}_1 > \bar{\beta}_2$. When this is the case, we refer to class 1 as the critical class and class 2 as the non-critical class. In this case, our proposed policy works as follows: whenever a critical order arrives, it is immediately satisfied if the on-hand inventory is positive, or backlogged if the on-hand inventory is zero. A non-critical order is accepted as it arrives. At its due date (T time units after its arrival), it is satisfied only if the on-hand inventory is above the critical level, S_c . If the inventory is at or below S_c , it is backlogged. Note again that, whether critical or non-critical, each demand arrival triggers a replenishment order which will arrive after L time units. Incoming replenishment orders are allocated according to a priority clearing mechanism. Under this mechanism, if there is a critical backorder at the time of a replenishment arrival it is immediately cleared, if there is a non-critical backorder it is cleared only if the on-hand inventory has reached S_c . In other words, incoming replenishment orders are used to clear backorders of the non-critical class only if the on-hand inventory is at the critical level, S_c . Given our rationing policy, the service level for the critical and non-critical classes clearly depend on S and S_c as well as parameters of the system: λ_1 , λ_2 , L , and T . We specifically assume backordering for both classes of demand as the company mentioned above is the primary (and most of the times the only) supplier of spare parts to its customers.

As we note in Section 1, there could be other applications which require $\bar{\beta}_2 > \bar{\beta}_1$. In that case, we refer to class 2 as the critical class and class 1 as the non-critical class. Our proposed policy works similarly. We only note that in this specific case, the critical level S_c is still static and does not depend on the number of class 2 orders collected (but not yet shipped). When $\bar{\beta}_1 = \bar{\beta}_2$, no rationing is applied

and we will later show that our model reduces to deterministic replenishment lead time version of the model in Wang et al. [23].

We also assume that $T \leq L$. This is a reasonable assumption since replenishment lead times are usually long and spare part providers cannot quote a demand lead time longer than the replenishment lead times. This assumption is also valid for the capital equipment manufacturer that motivated this research.

Given this system, our purpose is to determine the minimum inventory investment that satisfies the service requirements for both classes. Furthermore, we assume the ownership of on-order inventory and minimize expected inventory on hand plus expected inventory on order. Note that unlike the case in a standard continuous review $(S - 1, S)$ policy, the inventory position is not always equal to S in this system with demand lead times. The expected inventory position is in fact equal to $S + \lambda_2 \times T$, where the second term is due to the outstanding replenishment orders for the non-critical demand class that are yet not due. When we assume that fill rates are reasonably high, we can approximate the expected inventory on hand plus expected inventory on order by the expected inventory position. In Section 4.3, we show in a case study with 64 parts that the average backorder levels are very low with various high fill rates, verifying that $S + \lambda_2 \times T$ is a good approximation for the expected inventory on hand plus expected inventory on order. Thus we select our objective as minimizing S (since $\lambda_2 \times T$ is constant).

Observe that the service level for the critical class is closely related to the way incoming orders are handled and thus the arrival process. Therefore finding a closed form expression for the service level of the critical class is extremely difficult and for this reason we have to resort to approximations. In the next section, we will derive the service level of the non-critical class and an approximation for the service level of the critical class. We will then use the expressions for the service levels in an inventory optimization model in Section 3.2.

3.1 Deriving the Service Levels

In this section, we derive the resulting service levels for a given set of policy parameters S and S_c . The service level that we derive is exact for the non-critical demand class. The service level for the critical demand class, however, is an approximation. Later in this section, we will show analytically that the approximation constitutes a lower bound for the actual service level for the critical demand class, when we use a priority clearing mechanism to clear the backorders.

First, consider the service level for the non-critical demand class and consider the interval $(t, t + L]$. Since all outstanding orders at time t would arrive by time $t + L$, the inventory level at time $t + L$ would be S , if no demand occurred during the interval. In order for a non-critical demand that is

due at $t + L$ to be fulfilled at its due date, the inventory level at time $t + L$ must be at least $S_c + 1$ and this would happen if and only if the sum of the class 1 demand during $(t, t + L]$ and the class 2 demand due in $(t + T, t + L]$ is less than $S - S_c$. Observe that we do not need to consider the class 2 demand due in $(t, t + T]$ as the replenishments for these demands are already received by the time $t + L$, and hence, they do not impact the inventory level at time $t + L$. Thus, the service level of the non-critical demand class is given by

$$\beta_j^n(S, S_c) = P\{D_1(t, t + L] + D_2(t + T, t + L] \leq S - S_c - 1\}. \quad (1)$$

Letting $p(i; \lambda) = e^{-\lambda} \lambda^i / i!$, we have the following expression for the service level of the non-critical demand class

$$\beta_j^n(S, S_c) = \sum_{i=0}^{S-S_c-1} p(i; \lambda_1 L + \lambda_2(L - T)). \quad (2)$$

Now consider the service level for the critical demand and again consider the time interval $(t, t + L]$. Since all outstanding orders at time t would arrive by time $t + L$, the inventory level at time $t + L$ would be S , if no demand occurred during the interval. In order to satisfy a critical demand arriving at $t + L$, there must be at least one unit of inventory at $t + L$. Note that the replenishment orders corresponding to the class 2 demands that are due in the interval $(t, t + T]$ are received in the interval $(t + L - T, t + L]$. In order to calculate the probability that there is at least one unit of inventory at $t + L$, we condition on whether the hitting time, the arrival of the $S - S_c$ units of total demand that has a net impact on inventory, is in one of the two intervals $(t, t + L - T]$ or $(t + L - T, t + L]$, or after $t + L$.

$$\begin{aligned} \beta_j^c(S, S_c) &= P\{D_j(t + H, t + L] \leq S_c - 1, H \leq L - T\} \\ &+ P\{D_j(t + H, t + L] \leq S_c - 1, L - T \leq H \leq L\} + P\{H \geq L\}. \end{aligned} \quad (3)$$

In interval $(t, t + L - T]$, the density function of the hitting time can be found using

$$F_1(y) = P\{H \leq y\} = P\{D_1(t, t + y] + D_2(t, t + y] \geq S - S_c\}$$

In this region, the density $f_1(y) = dF_1(y)/dy$ of the hitting time can be derived as

$$f_1(y) = (\lambda_1 + \lambda_2)^{S-S_c} e^{-(\lambda_1 + \lambda_2)y} \frac{y^{S-S_c-1}}{(S - S_c - 1)!}$$

which is the density of the Erlang $S - S_c$ random variable with rate $\lambda_1 + \lambda_2$.

In interval $(t + L - T, t + L]$, the density function of the hitting time can be found using

$$F_2(y) = P\{H \leq y\} = P\{D_1(t, t + y] + D_2(t + y, t + L - T + y] \geq S - S_c\}.$$

Note that we are considering class 2 demands only in $(t + y, t + L - T + y]$, since the replenishment orders for class 2 demands in interval $(t, t + y]$ will be received by $t + L - T + y$. In this region, the density $f_2(y) = dF_2(y)/dy$ of the hitting time can be derived as

$$f_2(y) = \lambda_1 e^{-(\lambda_1 y + \lambda_2(L-T))} \frac{[\lambda_1 y + \lambda_2(L-T)]^{S-S_c-1}}{(S-S_c-1)!}$$

Finally, the hitting time is greater than or equal to L , if and only if the total net demand during the interval $(t, t + L)$ is less than $S - S_c$. Hence, we have

$$P\{H \geq L\} = P\{D_1(t, t + L] + D_2(t + T, t + L] \leq S - S_c - 1\} = \sum_{i=0}^{S-S_c-1} p(i; \lambda_1 L + \lambda_2(L-T)).$$

Thus the service level for the critical demand class can be written as

$$\begin{aligned} \beta_j^c(S, S_c) &= \int_0^{L-T} f_1(y) \times \left(\sum_{i=0}^{S_c-1} p(i; \lambda_j(L-y)) \right) dy + \int_{L-T}^L f_2(y) \times \left(\sum_{i=0}^{S_c-1} p(i; \lambda_j(L-y)) \right) dy \\ &+ \sum_{i=0}^{S-S_c-1} p(i; \lambda_1 L + \lambda_2(L-T)). \end{aligned} \quad (4)$$

The above expression is an approximation since it does not take into account how the incoming replenishment orders are handled after the hitting time. In fact, we next show that the expression is a lower bound for the actual service level when the incoming replenishment orders are handled according to a priority clearing mechanism.

Theorem 1 *The approximation for the critical service level given in Equation 3 is a lower bound for the actual critical service level, given that the priority clearing mechanism is employed, that is, all incoming replenishment orders are allocated to the critical class until the inventory on-hand reaches S_c .*

Proof: Since all outstanding replenishments at t will arrive at time $t + L$, we have the following

$$\begin{aligned} I(t) - B(t) + R(t, t + H] + R(t + H, t + L] &\geq S, \text{ or} \\ I(t) + R(t, t + H] &\geq S - R(t + H, t + L] + B(t). \end{aligned} \quad (5)$$

The inequality is due to the replenishment orders that correspond to the class 2 demands that arrive before $t + L$. In order to write the inventory level at time $t + H$, consider the worst case, i.e., no rationing has ever been performed during the interval $(t, t + H]$ and all backorders at time t are cleared by time $t + H$. Thus,

$$I(t + H) \geq I(t) + R(t, t + H] - D_1(t, t + H] - \hat{D}_2(t, t + H] - B(t). \quad (6)$$

where $\hat{D}_2(t, t+H)$ refers to the class 2 demands that have net impact on inventory. From Equations 5 and 6, we have

$$I(t+H) \geq S - R(t+H, t+L] - D_1(t, t+H] - \hat{D}_2(t, t+H].$$

However, by definition, $D_1(t, t+H] + \hat{D}_2(t, t+H] = S - S_c$. Therefore, we have

$$I(t+H) = S_c - R(t+H, t+L] + x, \text{ for some } x \geq 0.$$

The maximum level of inventory during the interval $(t+H, t+L]$ is $S_c + x$. Therefore, under a priority clearing mechanism, x is the maximum amount of inventory that could be used to satisfy non-critical demands or to clear non-critical backorders. Hence, we have

$$I(t+L) \geq I(t+H) + R(t+H, t+L] - D_j(t+H, t+L] - x, \text{ or,}$$

$$I(t+L) \geq S_c - D_j(t+H, t+L].$$

Since, we are conditioning on the event $\{D_j(t+H, t+L] \leq S_c - 1\}$, we have,

$$I(t+L) \geq 1.$$

□

The service level approximation for the critical class given in Equations 3 and 4 are valid for both $\bar{\beta}_1 > \bar{\beta}_2$ (class 1 is the critical class) and $\bar{\beta}_2 > \bar{\beta}_1$ (class 2 is the critical class).

The expressions for service level measures for the non-critical and critical demand classes given in Equations 2 and 4 are clearly linked to expressions that are developed in previous research. Note that we extend the single echelon model studied in Wang et al. [23] by introducing rationing to provide differentiated service for two demand classes (but we assume deterministic replenishment lead times). If we assume $S_c = 0$ in our model and deterministic replenishment lead times in Wang et al., we will see that the the service levels for the critical and non-critical demand classes can be both expressed as

$$\beta_j^n(S) = \beta_j^c(S) = P\{D_1(t, t+L] + D_2(t, t+L-T] \leq S-1\} = \sum_{i=0}^{S-1} p(i; \lambda_1 L + \lambda_2(L-T)) \quad (7)$$

The derivations are given in Appendix 1.

Note that our model extends the rationing models given in Dekker et al. [5] and Deshpande et al. [6] by introducing a demand lead time for the non-critical demand class. Dekker et al. study an $(S-1, S)$ inventory rationing model and derived an exact expression for the non-critical demand class and approximate expression for the critical demand class. If we assume $T = 0$ in our model, we will

see that the service levels for the critical and non-critical demand classes can be expressed as

$$\begin{aligned}\beta_j^c(S, S_c) &= \int_0^L (\lambda_1 + \lambda_2)^{S-S_c} e^{-(\lambda_1 + \lambda_2)y} \frac{y^{S-S_c-1}}{(S-S_c-1)!} \times \left(\sum_{i=0}^{S_c-1} p(i; \lambda_j(L-y)) \right) dy \\ &+ \sum_{i=0}^{S-S_c-1} p(i; (\lambda_1 + \lambda_2)L)\end{aligned}\quad (8)$$

$$\beta_j^n(S, S_c) = \sum_{i=0}^{S-S_c-1} p(i; (\lambda_1 + \lambda_2)L) \quad (9)$$

which are same as the expressions given in Dekker et al. [5] (with a slight change in notation).

In Deshpande et al. [6], the authors consider a (Q, r) inventory policy with non-critical demand backordered when the on-hand inventory falls below a threshold level K . When a replenishment order arrives existing backorders are cleared according to a special threshold clearing mechanism. Under this clearing mechanism, the backorders are cleared in the same manner as orders would be filled if there were more inventory available at the time demand arrived. In Appendix 2, we show that the service levels obtained through this “approximation” in Deshpande et al. [6] are exactly equal to the expressions given in Equations 8 and 9, if we assume $r = S - 1$ and $Q = 1$ and $K = S_c$ in their model.

Since our initial motivation was the case $\bar{\beta}_1 > \bar{\beta}_2$, we next develop simpler expressions and develop structural properties for $\beta_1^c(S, S_c)$. First, we show that equation 3 can be simplified if class 1 is the critical class. In this case, if the hitting time H is after $t + L - T$, say at time $t + L - T + z$, we need to consider class 2 demands that are due only in the interval $(t + z, t + L - T + z]$, as the replenishment orders corresponding to the class 2 demands that are due in period $(t, t + z]$ will arrive before $t + L - T + z$. If the hitting time H is after $t + L - T$, then all we care is whether the total demand in interval $(t, t + L]$ is less than $S - 1$. Thus, we can write the service level for class 1 as:

$$\begin{aligned}\beta_1^c(S, S_c) &= P\{D_1(t + H, t + L] \leq S_c - 1, H \leq L - T\} \\ &+ P\{D_1(t, t + L - T] + D_2(t, t + L - T] \leq S - S_c - 1, \\ &D_1(t, t + L] + D_2(t, t + L - T] \leq S - 1\}\end{aligned}\quad (10)$$

Hence, we do not need to specify the hitting time if it is later than $L - T$. Therefore, we have the following

$$\beta_1^c(S, S_c) = \int_0^{L-T} f_1(y) \times \left(\sum_{i=0}^{S_c-1} p(i; \lambda_1(L-y)) \right) dy + \sum_{i=0}^{S-S_c-1} \sum_{x=0}^{S-i-1} p(i; (\lambda_1 + \lambda_2)(L-T)) p(x; \lambda_1 T) \quad (11)$$

where $f_1(\cdot)$ is the density of the Erlang $S - S_c$ random variable with parameter $\lambda_1 + \lambda_2$.

An alternative way to derive the first part of the expression in (11) is to condition on the number of customer arrivals in $(t, t + L - T]$ and find out how many of the arrivals after $S - S_c$ are in fact

critical demands using Binomial distribution. Thus, we can write

$$\begin{aligned}\beta_1^c(S, S_c) &= \sum_{y=S-S_c}^{\infty} p(y; (\lambda_1 + \lambda_2)(L - T)) \sum_{x=0}^{S_c-1} p(x; \lambda_1 T) \sum_{i=0}^{S_c-1-x} b(\alpha_1, y - S + S_c, i) \\ &+ \sum_{i=0}^{S-S_c-1} \sum_{x=0}^{S-i-1} p(i; (\lambda_1 + \lambda_2)(L - T)) p(x; \lambda_1 T)\end{aligned}\quad (12)$$

where $b(\alpha_1, y - S + S_c, i)$ is the binomial probability with $\alpha_1 = \lambda_1/(\lambda_1 + \lambda_2)$ for $i \geq y - S + S_c$ and is equal to 0 for $i > y - S + S_c$.

We now develop structural properties for the approximation for the critical service level. We first note that the following lemma is immediately clear from Equations 2 and 4.

Lemma 1 $\beta_1^c(S, S_c) = \beta_2^n(S, S_c)$ if $S_c = 0$ and $\beta_1^c(S, S_c) \geq \beta_2^n(S, S_c)$ for all $S_c \geq 1$.

We next provide two lemmas which states that our approximation for the critical service level is monotone in the base stock level and critical level.

Lemma 2 *The approximation for the critical service level $\beta_1^c(S, S_c)$ given in Equation 12 is increasing in S .*

Proof: See Appendix 3.

Lemma 3 *The approximation for the critical service level $\beta_1^c(S, S_c)$ given in Equation 12 is increasing in S_c .*

Proof: See Appendix 4.

3.2 Service Level Optimization

We use the service levels for the critical and non-critical service levels that are derived in previous section to solve the following optimization problem to minimize inventory investment:

$$\min_{S, S_c} S \quad (13)$$

s.t.

$$\beta_1^c(S, S_c) \geq \bar{\beta}_1 \quad (14)$$

$$\beta_2^n(S, S_c) \geq \bar{\beta}_2 \quad (15)$$

$$S, S_c \geq 0 \quad (16)$$

The objective function in (13) is the base stock level. Note again that minimizing the base stock level S minimizes the expected inventory position $S + \lambda_2 \times T$ which approximates the expected on hand

plus on order inventory. The first constraint (14) states that the approximated service level for the critical demand class is higher than the minimum required service level, which ensures that the actual service level is also higher than the minimum required service level due to Theorem 1. The second constraint (15) ensures that the actual service level of the non-critical demand class is higher than the minimum required service level. The third constraint (16) ensures the non-negativity of the base stock and critical levels.

Table 1: Service level optimization algorithm

```

Set  $S_{min} := \arg \min \{x \geq 0 : \beta_2^n(x, 0) \geq \bar{\beta}_2\}$ 
Set  $S := S_{min}$ 
Set  $S_c := 0$ 
while  $\beta_1^c(S, S_c) < \bar{\beta}_1$  do
     $S := S + 1$ 
     $S_c := S - S_{min}$ 
end while
 $S^* = S$ 
 $S_c^* = S_c$ 
break

```

The algorithm for the optimization is presented in Table 1. It starts by determining S_{min} , the minimum amount of inventory needed to ensure $\bar{\beta}_2$ assuming all demand, if there was no rationing (i.e., $S_c = 0$). Note from equation 2 that the service level for the non-critical demand class depends only on the difference $S - S_c$, but not on S and S_c individually. Therefore in any solution that satisfies $\beta_2^n(S, S_c) \geq \bar{\beta}_2$, $S - S_c$ should be at least S_{min} . On the other hand, we show in Lemma 3 that the service level for the critical class $\beta_1^c(S, S_c)$ is increasing in S_c for a given S . Therefore, for a given S , it suffices to check whether $S_c = S - S_{min}$ satisfies the service level $\bar{\beta}_1$. Thus the algorithm starts with $S = S_{min}$ and S is increased at each iteration until the critical service level $\bar{\beta}_1$ is satisfied. It is also worthwhile to mention the maximum value of S that can be potentially searched in the algorithm. The maximum value of S is given by $S_{max} = \arg \min \{x \geq 0 : \beta_1^c(x, 0) \geq \bar{\beta}_1\}$, which is the minimum amount of inventory needed to ensure $\bar{\beta}_1$ (again assuming all demand), if there was no rationing. In other words, S_{max} is the solution of the simple round-up policy.

Since we know that our approximation for the critical service level is only a lower bound, there exists an opportunity to further reduce the base stock level found using the approximated critical service level in the optimization study. To do this, we conduct a simulation optimization where the service level for the critical demand class is derived through simulation. However we need to enumerate all possible (S, S_c) pairs (where $S \geq S_{min}$ and $S_c \geq S - S_{min}$) since we do not have the monotonicity results for the actual critical service level. The results of this study together with the output of the optimization study are provided in Section 4.2.

4 Numerical Study

Our numerical study is composed of three parts. In Section 4.1, we test the performance of the approximation for the critical service level that is suggested in Section 3.1 and identify the cases where we can estimate the actual service level with reasonable accuracy. To accomplish this, we use a simulation model coded in C and compare the simulated service level with service level calculated through the approximation. Having confirmed that the approximation works well in most cases, we use the approximation in the optimization model to demonstrate the impact of various factors on base stock levels and critical levels in Section 4.2. In Section 4.3 we demonstrate our results using a dataset from the capital equipment manufacturer that consists of 64 parts. Consistent with our motivation, in all of our experiments we have $\bar{\beta}_1 > \bar{\beta}_2$.

4.1 Simulation Study

In this section, we analyze the performance of the approximation for the critical service level with respect to the actual (simulated) service level. All tables in this section show the simulated non-critical service level, the exact non-critical service level calculated from Equation 2, the simulated critical service level, the approximation for the critical service level calculated from Equation 12, the difference between the simulated service level and the approximation for the critical service level and the percentage difference. The percentage difference is given by the percentage of the difference between the simulated critical service level and the approximation for the critical service level with respect to the simulated service level, that is, $100 \times (\text{simulation} - \text{approximation}) / \text{simulation}$.

4.1.1 Accuracy of the approximation for high service levels

First, we test the performance of the approximation when the required service level is high, specifically at 99 % and 95 %. Testing the approximation specifically at these levels is useful as high service levels are quite common in industry, especially for critical parts or critical demand classes. In Table 2, we test the performance of the approximation for the critical service level at around 99 % for 19 different instances. The supply lead time, L is 0.5 and the demand lead time, T is 0.1. The base stock level, the critical level and the arrival rates are chosen so that the resulting service level is around 99 %. First, note that for the non-critical demand class, the maximum difference between the service level obtained through the exact expression in Equation 2 and the service level obtained through simulation is 0.0005 (0.13 %), which shows that our simulation can accurately describe the system. Observe that the approximation works quite well when the critical service level is around 99 %. The average difference between the approximation and simulation is 0.0109 and the average percentage difference is 1.13 %. Note also that the approximation works better for higher service levels and in fact the best

Table 2: Performance of the approximation for a fixed SL of 99 % ($L=0.5$ and $T=0.1$)

λ_1	λ_2	S	S_c	β_2 (sim)	β_2 (exact)	β_1 (sim)	β_1 (approx)	Difference (sim-approx)	% difference
1	4	5	3	0.3791	0.3796	0.9995	0.9976	0.0019	0.19
2	4	6	3	0.5180	0.5184	0.9981	0.9927	0.0054	0.54
3	4	7	3	0.6249	0.6248	0.9968	0.9892	0.0076	0.76
4	4	8	3	0.7066	0.7064	0.9962	0.9877	0.0085	0.85
5	4	9	3	0.7693	0.7693	0.9958	0.9876	0.0082	0.82
6	4	10	3	0.8178	0.8180	0.9958	0.9884	0.0074	0.74
7	4	11	3	0.8561	0.8560	0.9960	0.9896	0.0064	0.64
8	4	12	3	0.8858	0.8857	0.9964	0.9909	0.0055	0.55
9	4	13	3	0.9091	0.9090	0.9967	0.9922	0.0045	0.45
10	4	14	3	0.9274	0.9274	0.9971	0.9934	0.0037	0.37
11	4	15	3	0.9421	0.9420	0.9975	0.9945	0.0030	0.30
12	4	16	3	0.9537	0.9536	0.9978	0.9954	0.0024	0.24
2	4	8	1	0.9828	0.9828	0.9983	0.9963	0.0020	0.20
3	4	8	2	0.9059	0.9057	0.9974	0.9928	0.0046	0.46
4	4	8	3	0.7066	0.7064	0.9962	0.9877	0.0085	0.85
5	4	8	4	0.4140	0.4142	0.9943	0.9802	0.0141	1.42
6	4	8	5	0.1623	0.1626	0.9923	0.9697	0.0226	2.28
7	4	8	6	0.0370	0.0372	0.9910	0.9554	0.0356	3.59
8	4	8	7	0.0037	0.0037	0.9921	0.9368	0.0553	5.57

performance is achieved for the case when the service level is highest. This is because at high service levels for the critical demand class, the backorders primarily consist of backorders for the non-critical demand class, and the way incoming replenishment orders -which is the major shortcoming of the approximation- is less important.

We repeat the analysis above for a critical service level around 95 % in Table 3 for 10 different instances. Again, the supply lead time, L is 0.5 and the demand lead time, T is 0.1. The approximation still works fine, although the performance is not as good as the 99 % service level case. The average difference between the approximation and simulation is 0.0134 and the average percentage difference is 1.43 %. Note again that the approximation works better for higher service levels and the best performance is achieved for cases when the service level is highest around 97 %.

Table 3: Performance of the approximation for a fixed SL of 95 % ($L=0.5$ and $T=0.1$)

λ_1	λ_2	S	S_c	β_2 (sim)	β_2 (exact)	β_1 (sim)	β_1 (approx)	Difference (sim-approx)	% difference
4	1	5	2	0.5696	0.5697	0.9380	0.9190	0.0190	2.03
5	1	6	2	0.6705	0.6696	0.9481	0.9339	0.0142	1.50
6	1	7	2	0.7440	0.7442	0.9573	0.9467	0.0106	1.11
7	1	8	2	0.8000	0.8006	0.9652	0.9573	0.0079	0.82
8	1	9	2	0.8433	0.8436	0.9718	0.9658	0.0060	0.62
9	1	10	2	0.8765	0.8769	0.9772	0.9726	0.0046	0.47
5	1	7	1	0.9262	0.9258	0.9761	0.9722	0.0039	0.40
6	1	7	2	0.7440	0.7442	0.9573	0.9467	0.0106	1.11
7	1	7	3	0.4535	0.4532	0.9321	0.9118	0.0203	2.18
8	1	7	4	0.1855	0.1851	0.9040	0.8671	0.0369	4.08

4.1.2 Accuracy of the approximation with varying system parameters

In Table 4, we allow the critical service level to vary and we test the performance of the approximation by varying a single parameter such as base stock level, arrival rate for the critical demand class, arrival rate for the non-critical demand class and demand lead time. As seen from the first part of the table, critical and non-critical service levels both increase as the base stock level increases. We also note that

Table 4: Performance of the approximation with varying system parameters

S	S_c	λ_1	λ_2	L	T	β_2 (sim)	β_2 (exact)	β_1 (sim)	β_1 (approx)	Diff (sim-approx)	% diff
7	2	6	2	0.50	0.10	0.6684	0.6678	0.9486	0.9225	0.0261	2.75
8	2	6	2	0.50	0.10	0.8160	0.8156	0.9773	0.9655	0.0118	1.21
9	2	6	2	0.50	0.10	0.9094	0.9091	0.9909	0.9861	0.0048	0.48
10	2	6	2	0.50	0.10	0.9600	0.9599	0.9967	0.9949	0.0018	0.18
11	2	6	2	0.50	0.10	0.9840	0.9840	0.9989	0.9983	0.0006	0.06
5	2	1	1	1.00	0.50	0.8090	0.8088	0.9950	0.9860	0.0090	0.90
5	2	2	1	1.00	0.50	0.5439	0.5438	0.9481	0.9008	0.0473	4.99
5	2	3	1	1.00	0.50	0.3218	0.3208	0.8377	0.7378	0.0999	11.93
5	2	4	1	1.00	0.50	0.1739	0.1736	0.6961	0.5438	0.1523	21.88
5	2	5	1	1.00	0.50	0.0886	0.0884	0.5614	0.3668	0.1946	34.66
5	2	1	1	1.00	0.50	0.8090	0.8088	0.9950	0.9860	0.0090	0.90
5	2	1	2	1.00	0.50	0.6770	0.6767	0.9936	0.9686	0.0250	2.52
5	2	1	3	1.00	0.50	0.5436	0.5438	0.9928	0.9484	0.0444	4.47
5	2	1	4	1.00	0.50	0.4227	0.4232	0.9923	0.9274	0.0649	6.54
5	2	1	5	1.00	0.50	0.3207	0.3208	0.9921	0.9072	0.0849	8.56
14	3	10	4	0.50	0.10	0.9274	0.9274	0.9971	0.9934	0.0037	0.37
14	3	10	4	0.50	0.20	0.9486	0.9486	0.9983	0.9953	0.0030	0.30
14	3	10	4	0.50	0.30	0.9651	0.9651	0.9990	0.9973	0.0017	0.17
14	3	10	4	0.50	0.40	0.9775	0.9775	0.9994	0.9986	0.0008	0.08
14	3	10	4	0.50	0.50	0.9946	0.9945	0.9995	0.9993	0.0002	0.02

the difference between actual and approximated service level decreases confirming the performance of our approximation for high critical service levels. In the second and third part of the table, we study the impact of the critical arrival rate and the non-critical arrival rates, respectively. As we increase both rates, we see that both critical and non-critical service levels deteriorate. As we already observed before, the performance of the approximation also deteriorates as we begin to see low service levels. The difference between the simulated and approximated critical service levels are at unacceptable levels for service levels around 60 %. However, note that these service levels are hardly observed in practice, especially for critical items or for critical demand classes. In the fourth part of Table 4, we study the impact of demand lead time, T . As T increases, both the critical and non-critical service levels. Again, the difference behaves as expected, attaining its smallest value when the critical service level is the highest. As expected, we also see that the non-critical service level is quite sensitive to the demand lead time, while the critical service level is not.

The results in our simulation study show that, with a reasonable accuracy, our approximation can be used to estimate the actual service levels for the critical demand class when a priority clearing mechanism is used. In all of our experiments, service level obtained through approximation is lower than the actual service level for the critical demand class as proven in Section 3. Finally, we observe that the performance of the approximation improves as the service level for the critical demand class increases which is in line with high service level needs for critical demand classes.

4.2 Optimization Study

In this section, we present the output of our optimization and simulation optimization study to demonstrate that a system with rationing, even when the approximation for the critical service level is used, can result in significant inventory savings compared to one without rationing. Through Tables 5-7

Table 5: Optimal Parameters: Approximation vs Simulation ($\lambda_1 = 1$, $L=0.5$, $T=0.1$, $\bar{\beta}_1 = 0.99$ and $\bar{\beta}_2 = 0.80$)

λ_2	Round-up S	Approximation			Simulation					
		S	S_c	% saving	S^*	S_c^*	% saving	β_1	$\beta_1(\text{app})$	β_2
1	5	4	1	20.00	4	1	20.00	0.9933	0.9903	0.9372
2	6	5	2	16.67	5	2	16.67	0.9988	0.9972	0.8570
3	6	6	0	0.00	5	1	16.67	0.9926	0.9852	0.9067
4	7	6	2	14.29	6	2	14.29	0.9992	0.9971	0.8386
5	8	7	2	12.50	6	1	25.00	0.9930	0.9830	0.8912
6	8	7	2	12.50	7	2	12.50	0.9994	0.9972	0.8317
7	9	8	2	11.11	7	1	22.22	0.9937	0.9820	0.8828
8	10	8	2	20.00	7	1	30.00	0.9908	0.9731	0.8301
9	10	9	2	10.00	8	1	20.00	0.9943	0.9817	0.8786
10	11	9	2	18.18	8	1	27.27	0.9921	0.9739	0.8310

we present our results for various input parameters. In all tables, the first part represents the input parameter in consideration. The second part shows the required base stock level if a round-up policy is used (demand lead times are still recognized). This base stock level is determined by the higher service level requirement which is the service level requirement for class 1 in our case. The third part shows the results with rationing and when the approximation is used for the critical service level. The fourth part shows the results with rationing and when simulation is used for the critical service level. For the last two parts, we report the percentage savings in the base stock levels that are obtained with rationing against a round-up policy.

In Table 5, λ_2 varies between 1 and 10 while λ_1 is kept fixed at 1. We can reach the optimal solution for four cases using our approximation, whereas in other cases there is only a single unit gap. The fine performance of the optimization study that uses approximation is attributed to the relatively small arrival rates and lead time demands. In addition, observe that rationing tends to create more savings when the arrival rate in the non-critical demand class is significantly higher than the arrival rate in the critical demand class, although there is no uniformity in this behavior.

In Table 6, $\bar{\beta}_1$ varies between 90 percent and 99.5 percent, while $\bar{\beta}_2$ is kept fixed at 80 percent. It can easily be seen that rationing becomes more effective as the critical service level requirement increases.

Table 6: Optimal Parameters: Approximation vs Simulation ($\lambda_1 = 5$, $\lambda_2 = 10$, $L=2$, $T = 0.5$ and $\bar{\beta}_2 = 0.80$)

$\bar{\beta}_1$	Round-up S	Approximation			Simulation					
		S	S_c	% saving	S^*	S_c^*	% saving	β_1	$\beta_1(\text{app})$	β_2
0.900	33	32	2	3.03	31	1	6.06	0.9544	0.8179	0.8650
0.925	33	33	0	0.00	31	1	6.06	0.9544	0.8179	0.8650
0.950	34	34	0	0.00	31	1	8.82	0.9544	0.8179	0.8650
0.970	36	35	5	2.78	32	2	11.11	0.9889	0.9061	0.8179
0.980	37	35	5	5.41	32	2	13.51	0.9889	0.9061	0.8179
0.985	37	36	6	2.70	32	2	13.51	0.9889	0.9061	0.8179
0.990	38	36	6	5.26	33	3	13.16	0.9973	0.9401	0.8179
0.995	40	37	7	7.50	33	3	17.50	0.9973	0.9401	0.8179

The results in Tables 5 and 6 are in line with the numerical results that are found in Deshpande

et al. [6]. Namely, rationing is more effective in cases where the non-critical arrivals are dominant in the arrival mix (λ_2 is large compared to λ_1) and when the service level requirements are significantly different ($\bar{\beta}_1$ is large as compared to $\bar{\beta}_2$).

We next study the impact of demand lead time in Tables 7 and 8. First note that as the demand lead time T increases, all base stock parameters decrease. This is expected since as Hariharan and Zipkin [12] noted, demand lead times work just as the opposite of supply lead times. The impact of the demand lead time on the relative effectiveness of rationing (performance of the simulation against the round-up policy) does not have any functional form as we see from Tables 7 and 8, and numerous other experiments that we conducted. Savings are possible when the demand lead time is small or large, the extent of savings depends on the relative arrival rate of the class 2 customers, rather than the demand lead time of this class. It is important to note one interesting case in Tables 7 and 8. Even when $T = L$, rationing can still help reduce the base stock levels. In fact, in the second portion of Table 8, the maximum savings are realized when $T = L = 0.5$. This case is interesting, because if $T = L$, a separate stock policy would not even stock any items for class 2 customers and a round-up policy would not save any inventory. Even in this situation, pooling class 2 customers with class 1 customers and incorporating rationing will reduce the inventory levels, since a rationing policy can effectively utilize replenishments for class 2 demands to support class 1 service level.

In order to understand the impact of T on the performance of the approximation, remember that we consider the interval $(t, t + L]$ in the development of the approximation. As T approaches L , there are more class 2 demands whose replenishments are received before $t + L$. Since the approximation fails to treat the portion of such replenishments that are received after the hitting time, the approximation's performance degrades and fails to capture the savings as T increases.

Table 7: Optimal Parameters: Approximation vs Simulation ($\lambda_1 = 5$, $\lambda_2 = 10$, $\bar{\beta}_1 = 0.99$ and $\bar{\beta}_2 = 0.80$)

L	T	Round-up		Approximation			Simulation					
		S	$\beta_1 = \beta_2$	S	S_c	% saving	S^*	S_c^*	% saving	β_1	$\beta_1(\text{app})$	β_2
2.0	0.2	42	0.9920	38	5	9.52	36	3	14.29	0.9968	0.9590	0.8041
2.0	0.4	40	0.9936	37	6	7.50	34	3	15.00	0.9972	0.9438	0.8136
2.0	0.6	37	0.9918	36	7	2.70	32	3	13.51	0.9973	0.9389	0.8219
2.0	0.8	35	0.9936	34	7	2.86	29	2	17.14	0.9903	0.9136	0.8324
2.0	1.0	32	0.9919	32	0	0.00	27	2	15.63	0.9909	0.9222	0.8415
2.0	1.2	30	0.9941	30	0	0.00	25	2	16.67	0.9924	0.9318	0.8550
2.0	1.4	27	0.9925	27	0	0.00	22	2	18.52	0.9901	0.9108	0.8121
2.0	1.6	24	0.9907	24	0	0.00	20	2	16.67	0.9914	0.9235	0.8271
2.0	1.8	22	0.9939	22	0	0.00	18	2	18.18	0.9925	0.9370	0.8444
2.0	2.0	19	0.9928	19	0	0.00	17	3	10.53	0.9971	0.9730	0.9165
0.5	0.05	15	0.9943	13	3	13.33	12	2	20.00	0.9941	0.9800	0.8304
0.5	0.10	14	0.9929	13	3	7.14	12	2	14.29	0.9962	0.9824	0.8773
0.5	0.15	13	0.9912	12	3	7.69	11	2	15.38	0.9953	0.9723	0.8472
0.5	0.20	13	0.9955	12	4	7.69	10	2	23.08	0.9940	0.9593	0.8095
0.5	0.25	12	0.9945	11	3	8.33	10	2	16.67	0.9965	0.9724	0.8666
0.5	0.30	11	0.9933	11	0	0.00	9	2	18.18	0.9955	0.9628	0.8310
0.5	0.35	10	0.9919	10	0	0.00	9	2	10.00	0.9975	0.9791	0.8893
0.5	0.40	9	0.9901	9	0	0.00	8	2	11.11	0.9965	0.9734	0.8576
0.5	0.45	9	0.9962	9	0	0.00	7	2	22.22	0.9939	0.9665	0.8153
0.5	0.50	8	0.9958	8	0	0.00	7	2	12.50	0.9947	0.9858	0.9580

Table 8: Optimal Parameters: Approximation vs Simulation ($\lambda_1 = 1$, $\lambda_2 = 5$, $\bar{\beta}_1 = 0.99$ and $\bar{\beta}_2 = 0.90$)

L	T	Round-up		Approximation			Simulation					
		S	$\beta_1 = \beta_2$	S	S_c	% saving	S^*	S_c^*	% saving	β_1	$\beta_1(\text{app})$	β_2
2	0.2	20	0.9907	18	2	10.00	17	1	15.00	0.9911	0.9056	0.9719
2	0.4	19	0.9928	17	2	10.53	16	1	15.79	0.9920	0.9154	0.9684
2	0.6	18	0.9947	16	2	11.11	15	1	16.67	0.9933	0.9253	0.9685
2	0.8	16	0.9918	15	2	6.25	14	1	12.50	0.9943	0.9357	0.9711
2	1.0	15	0.9943	14	3	6.67	12	1	20.00	0.9913	0.9008	0.9521
2	1.2	13	0.9912	13	0	0.00	11	1	15.38	0.9930	0.9159	0.9598
2	1.4	12	0.9945	12	0	0.00	10	1	16.67	0.9950	0.9319	0.9689
2	1.6	10	0.9919	10	0	0.00	9	1	10.00	0.9965	0.9490	0.9788
2	1.8	9	0.9962	9	0	0.00	7	1	22.22	0.9944	0.9165	0.9665
2	2.0	7	0.9955	7	0	0.00	6	1	14.29	0.9935	0.9837	0.9834
0.5	0.05	8	0.9927	7	1	12.50	7	1	12.50	0.9964	0.9392	0.9927
0.5	0.10	8	0.9958	7	1	12.50	7	1	12.50	0.9977	0.9582	0.9940
0.5	0.15	7	0.9916	7	0	0.00	6	1	14.29	0.9957	0.9219	0.9864
0.5	0.20	7	0.9955	6	1	14.29	6	1	14.29	0.9973	0.9476	0.9902
0.5	0.25	6	0.9909	6	0	0.00	6	0	0.00	0.9909	0.9909	0.9909
0.5	0.30	6	0.9955	6	0	0.00	5	1	16.67	0.9971	0.9346	0.9860
0.5	0.35	5	0.9909	5	0	0.00	5	0	0.00	0.9909	0.9909	0.9909
0.5	0.40	5	0.9963	5	0	0.00	4	1	20.00	0.9967	0.9203	0.9828
0.5	0.45	4	0.9927	4	0	0.00	4	0	0.00	0.9927	0.9927	0.9927
0.5	0.50	4	0.9982	4	0	0.00	3	1	25.00	0.9916	0.9865	0.9856

Regarding the performance of our approximation, we note that the optimization algorithm that uses the approximation can reach the optimal values for cases where the arrival rates and thus lead time demands are relatively low (slow moving items). In these specific cases, the optimal base stock held to ensure the required service levels is low and there is a relatively small opportunity to save from rationing (one or two units) and our optimization algorithm can capture those savings. However, as the demand rates increase, more inventory is held to ensure the required service levels and hence there is more opportunity to obtain inventory savings from rationing. In these cases, our optimization algorithm with approximation usually cannot obtain the optimal base stock levels as it fails to accurately capture the important effect of incoming replenishment orders. Despite this fact, our optimization algorithm with approximation usually misses the optimal base stock level by only a few units. We have to note at this point that our optimization algorithm which uses the approximation for the critical service level is significantly faster than the simulation optimization. On a 2 Ghz Pentium 4 processor, the run time of the optimization algorithm is 5 minutes on the average, while the average run time for a single simulation is 60 minutes. Because the simulation optimization runs several times to find the actual critical service levels for different (S, S_c) pairs, this duration can increase to several hours depending on system parameters. In addition, observe that the results of our optimization algorithm with approximation are still significant even if one wants to use a simulation optimization based approach: the base stock levels that are obtained through approximation can be used as an upper bound which will reduce the computational burden considerably.

4.3 Case Study

In this section, we will show the significance of our results through a case study at the capital equipment manufacturer that we briefly described in Section 1. We have selected a depot in North America that is serving a number of customers for both down demand and lead time demand.

We have selected 64 parts for our study. A summary of the characteristics of these parts are given in Table 9. In order to ensure the appropriateness of the $(S - 1, S)$ inventory policy and the validity of the Poisson demand assumption, we include rather expensive and slow moving parts in our study. We used the demand history of a 12 month period in years 2001 and 2002 and include all requested orders (these could include orders that were not satisfied or canceled later), for which the primary source is the depot we have selected. The ratio of critical orders to total orders vary at the part level. On the average, 52.2 % of a part's demand is from down orders (i.e., the critical demand). In the same 12 month period, these 64 parts had a sales volume of \$ 41.2 million (in cost). \$ 24.3 million (59.1 %) of this is generated by down orders; \$ 16.9 million (40.9 %) of this is generated by lead time orders.

Table 9: Part Characteristics

	Min	Max	Average
Part Cost (\$)	1,104	40,451	8,681
Critical Annual Demand	1	166	43.19
Non-critical Annual Demand	2	120	39.59
Total Annual Demand	41	212	82.78
Percentage of Critical Demand	1.18	96.77	52.20
COGS (\$)	94,985	3,318,438	643,600
Replenishment Lead Time (days)	19	120	68.06
Lead Time Demand	10.06	19.65	13.99

The analysis is done in three steps. In the first step, we do not recognize the demand lead time for lead time orders and we do not apply any rationing. We simply calculate the minimum base stock levels that will satisfy the target service level requirement for the down orders considering the total demand (down demand plus lead time demand). This reflects the current practice in the company. Note also that in this current practice, the company initiates replenishment orders for lead time orders when the orders are physically shipped from the depot, as opposed to a policy that recognizes demand lead times which triggers a replenishment order as soon as the customer order is received. Therefore, the base stock level is exactly equal to inventory on hand plus inventory on order minus backorders with company's current practice.

In the second step, we recognize the demand lead time for lead time orders, however do not use any rationing to provide differentiated service to the two type of demand classes. We calculate the minimum base stock levels that will satisfy the target service level requirement for the down orders. This is similar to the model in [23] and in fact lead time orders and down orders get the same service level in this case. Finally, in the third step, we also use rationing to provide differentiated service to

two demand classes. In this analysis, we use the approximation for the critical service level that is derived in Section 3.1.

Table 10 shows the dollar value of base stock levels, total average inventory (on hand plus on order), average on hand inventory and average backorders (in million USD) across all parts for three different approaches. We see that recognizing the demand lead times and using rationing to differentiate service levels generate significant savings to the company for these 64 parts. For example, when the critical service level is 99 % and the non-critical service level is 80 %, recognizing demand lead times saves 2.41 % on total inventory (which is equal to the inventory investment, once we assume that the pipeline stocks are owned by the company); additional 3.65 % is saved once the company starts rationing (even though we use an approximation for the service level of the critical demand class) to provide differentiated services to two types of demand. We also note that if the company were responsible for only on hand inventory, the savings would be more significant. As the critical service level declines to approach the non-critical service level, we see that savings due to the recognition of demand lead times are still significant, while the impact of rationing is less pronounced.

Table 10: Impact of Critical Service Level

Critical SL (%)	90	95	97	99
Non Critical SL (%)	80	80	80	80
No Demand Lead Time - No Rationing				
Base Stock	11.449	12.294	12.930	14.050
Total Inventory	11.517	12.324	12.946	14.054
Inventory On Hand	3.387	4.195	4.816	5.924
Backlog	0.069	0.030	0.015	0.004
No Rationing				
Base Stock	10.636	11.450	12.007	13.009
Total Inventory	11.402	12.179	12.724	13.716
Savings (%)	1.00	1.18	1.71	2.41
Inventory On Hand	3.272	4.049	4.594	5.586
Savings (%)	3.41	3.46	4.59	5.72
Backlog	0.064	0.028	0.015	0.005
Rationing (Approx)				
Base Stock	10.563	11.233	11.652	12.432
Total Inventory	11.341	11.997	12.415	13.202
Savings (%)	1.53	2.65	4.10	6.06
Inventory On Hand	3.211	3.867	4.285	5.072
Savings (%)	5.20	7.80	11.02	14.38
Backlog	0.076	0.063	0.061	0.068

One other interesting observation about the analysis is that the average level of backlogs are very insignificant (\$ 76,000 in the worst case, \$ 4,000 in the best case) as compared to the base stock levels and cost of goods sold (\$ 41.2 million for all parts). This justifies our approach to minimize base stock levels in our model.

We conclude that the recognition of the demand lead times and the use of rationing create significant savings for the company. This is true even when we use an approximation to estimate the service level for the critical demand class. More savings are obviously possible if we can accurately determine the service level for the critical demand class. However, the approximation is easy to implement (which is

necessary for this particular company) and as it is shown here, its performance is quite reasonable.

5 Conclusion

In this study, we consider a single echelon spare part distribution system. Our research is motivated by our experience with a leading capital equipment manufacturer. This manufacturer faces two kinds of orders: down orders that result from the equipment failures of the customers and lead time orders that result from the scheduled maintenance activities of customers. The down orders must be supplied immediately while the lead time orders are needed to be supplied at a future date (demand lead time). Currently, the company uses a common pool of inventory controlled by a base stock policy. However, it neither recognizes the demand lead times for the lead time demand, nor treats the down demand and lead time demand differently. Since unused manufacturing capacity in customers is very costly, we need a policy that could favor down orders. Therefore, we model the system as a single-echelon inventory model where down orders are considered as the critical demand class, while the lead time orders are considered as the non-critical class. For this setting, we propose a static rationing policy that would ration the demand from non-critical class. The model aims to satisfy the minimum service level requirements for both demand classes, while using less inventory than the policy the company currently uses.

We develop an approximation for the critical service level and prove that this approximation is essentially a lower bound for the critical service level. Furthermore we conduct a simulation study to test the performance of our approximation versus the actual (simulated) critical service level and show that our approximation performs quite well specifically for high critical service levels which is in line with the needs of a critical demand class. As a result of our optimization study, we show that rationing the non-critical orders indeed results in significant savings in terms of base stock inventory. Savings are also verified in a case study where we use real data from the capital equipment manufacturer we mentioned earlier.

We also present the situations where the rationing policy is most useful. The rationing policy is more effective when the non-critical arrival rate dominates the critical arrival rate and the critical service level requirement dominates the non-critical service level requirement. Obviously more savings are possible through using the actual service levels from simulation. However, our optimization algorithm which is extremely faster than the simulation optimization captures most of the savings due to rationing and hence, will be more effective especially in systems with many parts (as in the capital equipment manufacturer we consider).

During this research, we combine the concept of demand lead time and inventory rationing, both of which have proven to be cost effective for inventory systems. However, to our knowledge our study

is the first to simultaneously consider rationing and demand lead time. Our numerical results indicate that such a practice indeed results in significant inventory and cost savings.

Future research can extend the analysis here in many directions. Although we were motivated by a system with two demand classes, considering the possibility of more demand classes would be a useful extension. In addition, considering several deterministic demand lead times as well as stochastic supply or demand lead time is a logical direction. Obviously most inventory systems have a multi-echelon structure and therefore extending the analysis to a multi-echelon setting is another research avenue.

Appendix

Appendix 1.

It is trivial to derive equation 7 from equations 2 and 4 when $S_c = 0$. In Wang et al. [23], the service levels are given in Equations 26 and 27 which we repeat below:

$$\begin{aligned}\beta_1 &= P\{S - Q(0) + R(0) \geq 1\} \\ \beta_2 &= P\{S - Q(0) + R(0) \geq 1\} + G_2(T)P\{S - Q(0) + R(0) = 0\}\end{aligned}$$

where $Q(0)$ and $R(0)$ are Poisson random variables with mean values of $\lambda_1 q_1(0) + \lambda_2 q_2(0)$ and $\lambda_1 r_1(0) + \lambda_2 r_2(0)$, respectively, and

$$q_1(0) = \int_0^\infty [1 - G_1(x)]dx, \quad q_2(0) = \int_T^\infty [1 - G_2(x)]dx, \quad r_1(0) = \int_0^0 G_1(x)dx, \quad r_2(0) = \int_0^T G_2(x)dx.$$

Here, $G_1(\cdot)$ and $G_2(\cdot)$ are the cumulative distribution functions for replenishment lead times of critical and non-critical demand classes, respectively. Note that in our model we assume the lead times are deterministic at L and $T \leq L$. Thus, we have the following:

$$q_1(0) = L, \quad q_2(0) = L - T, \quad r_1(0) = 0, \quad r_2(0) = 0.$$

Thus, $Q(0)$ is a Poisson random variable with mean value $\lambda_1 L + \lambda_2(L - T)$ and R_0 is a Poisson random variable with mean value of 0. Also note that $G_2(T) = 0$ since $T \leq L$. Hence,

$$\beta_1 = \beta_2 = \sum_{i=0}^{S-1} p(i; \lambda_1 L + \lambda_2(L - T))$$

Appendix 2.

In Deshpande et al. [6], the service levels are given in Equations 13 and 14, which we report below

$$\begin{aligned}\beta_1(Q, r, K) &= 1 - \frac{1}{Q} \sum_{y=r+1}^{r+Q} a_1(y, K) \\ \beta_2(Q, r, K) &= 1 - \frac{1}{Q} \sum_{y=r+1}^{r+Q} a_2(y, K)\end{aligned}$$

where

$$\begin{aligned}
a_1(y, K) &= \sum_{x=y}^{\infty} \sum_{j=0}^{x-y} b(\alpha_1, x - y + K, K + j) p(x; \lambda L) \\
a_2(y, K) &= \begin{cases} \sum_{x=y-K}^{\infty} p(x; \lambda L) & \text{if } K \leq y, \\ 1 & \text{if } K > y. \end{cases}
\end{aligned}$$

Here $\lambda = \lambda_1 + \lambda_2$ and $b(\alpha_1, x - y + K, K + j)$ is the binomial probability with $\alpha_1 = \lambda_1/\lambda$. First, note that $r + 1 = S$, $Q = 1$ and $K = S_c < r + 1 = S$ in our model. Thus, the service level for the non-critical demand class reduces to

$$\beta_2 = 1 - \sum_{y=S}^{\infty} \sum_{x=y-S_c}^{\infty} p(x; \lambda L) = 1 - \sum_{x=S-S_c}^{\infty} p(x; \lambda L) = \sum_{x=0}^{S-S_c-1} p(x; \lambda L) \quad (17)$$

which is same as our expression for the non-critical demand class if we assume $T = 0$.

When $r + 1 = S$, $Q = 1$ and $K = S_c < r + 1 = S$ the service level for the critical demand class reduces to

$$\begin{aligned}
\beta_1 &= 1 - \sum_{y=S}^{\infty} \sum_{x=y}^{\infty} \sum_{j=0}^{x-y} b(\alpha_1, x - S + S_c, S_c + j) p(x; \lambda L) \\
&= 1 - \sum_{x=S}^{\infty} p(x; \lambda L) \sum_{j=0}^{x-S} b(\alpha_1, x - S + S_c, S_c + j) \\
&= 1 - \sum_{x=S}^{\infty} p(x; \lambda L) \left[1 - \sum_{j=0}^{S_c-1} b(\alpha_1, x - S + S_c, j) \right] \\
&= 1 - \sum_{x=S}^{\infty} p(x; \lambda L) + \sum_{x=S}^{\infty} p(x; \lambda L) \sum_{j=0}^{S_c-1} b(\alpha_1, x - S + S_c, j) \\
&= \sum_{x=0}^{S-1} p(x; \lambda L) + \sum_{x=S}^{\infty} p(x; \lambda L) \sum_{j=0}^{S_c-1} b(\alpha_1, x - S + S_c, j)
\end{aligned}$$

Now consider our approximation for the critical service level when $T = 0$. When $T = 0$, Equation 3 reduces to

$$\beta_1^c(S, S_c) = P\{D_1(t, t + L] + D_2(t, t + L] \leq S - S_c - 1\} + P\{D_1(t + H, t + L] \leq S_c - 1, H \leq L\}.$$

The first part of the sum can be derived using the same approach in Deshpande et al. [6]. For a given number of x total demand during the lead-time, $x - S + S_c$ will arrive after the hitting time. Out of these, the probability of having j critical demands is given by $b(\alpha_1; x - S + S_c; j)$. Thus, we have,

$$\beta_1^c(S, S_c) = \sum_{x=0}^{S-S_c-1} p(x; \lambda L) + \sum_{x=S-S_c}^{\infty} p(x; \lambda L) \sum_{j=0}^{S_c-1} b(\alpha_1; x - S + S_c; j).$$

The above argument can be rewritten as

$$\begin{aligned}
\beta_1^c(S, S_c) &= \sum_{x=0}^{S-S_c-1} p(x; \lambda L) + \sum_{x=S-S_c}^{S-1} p(x; \lambda L) \sum_{j=0}^{S_c-1} b(\alpha_1, x - S + S_c, j) \\
&\quad + \sum_{x=S}^{\infty} p(x; \lambda L) \sum_{j=0}^{S_c-1} b(\alpha_1, x - S + S_c, j) \\
&= \sum_{x=0}^{S-S_c-1} p(x; \lambda L) + \sum_{x=S-S_c}^{S-1} p(x; \lambda L) + \sum_{x=S}^{\infty} p(x; \lambda L) \sum_{j=0}^{S_c-1} b(\alpha_1, x - S + S_c, j) \\
&= \sum_{x=0}^{S-1} p(x; \lambda L) + \sum_{x=S}^{\infty} p(x; \lambda L) \sum_{j=0}^{S_c-1} b(\alpha_1, x - S + S_c, j)
\end{aligned}$$

Hence, the service level expression for the critical demand class when the threshold clearing mechanism is used is same as our approximation for the service level for the critical demand class.

Appendix 3: Proof of Lemma 1

Let $\Delta_S = \beta_1^c(S+1, S_c) - \beta_1^c(S, S_c)$. Also denote $p_1(y) = p(y; (\lambda_1 + \lambda_2)(L - T))$ and $p_2(x) = p(x; \lambda_1 T)$.

Then, we have

$$\begin{aligned}
\Delta_S &= \sum_{y=S+1-S_c}^{\infty} p_1(y) \sum_{x=0}^{S_c-1} p_2(x) \sum_{i=0}^{S_c-1-x} b(\alpha_1, y - S + S_c - 1, i) \\
&\quad - \sum_{y=S-S_c}^{\infty} p_1(y) \sum_{x=0}^{S_c-1} p_2(x) \sum_{i=0}^{S_c-1-x} b(\alpha_1, y - S + S_c, i) \\
&\quad + \sum_{i=0}^{S-S_c} \sum_{x=0}^{S-i} p_1(i) p_2(x) - \sum_{i=0}^{S-S_c-1} \sum_{x=0}^{S-i-1} p_1(i) p_2(x) \\
&= \sum_{y=S+1-S_c}^{\infty} p_1(y) \sum_{x=0}^{S_c-1} p_2(x) \left[\sum_{i=0}^{S_c-1-x} b(\alpha_1, y - S + S_c - 1, i) - \sum_{i=0}^{S_c-1-x} b(\alpha_1, y - S + S_c, i) \right] \\
&\quad - p_1(S - S_c) \sum_{x=0}^{S_c-1} p_2(x) \sum_{i=0}^{S_c-1-x} b(\alpha_1, 0, i) \\
&\quad + \sum_{i=0}^{S-S_c-1} p_1(i) p_2(S - i) + p_1(S - S_c) \sum_{x=0}^{S_c} p_2(x)
\end{aligned}$$

Note that

$$\sum_{i=0}^k b(\alpha, N-1, i) - \sum_{i=0}^k b(\alpha, N, i) = \begin{cases} \alpha b(\alpha, N-1, k), & \text{if } k < N-1 \\ 1 - \sum_{i=0}^{N-1} b(\alpha, N, i), & \text{if } k = N-1 \\ 0, & \text{if } k > N-1 \end{cases}$$

The last two cases can also be written as $\alpha b(\alpha, N-1, k)$ as we have

$$1 - \sum_{i=0}^{N-1} b(\alpha, N, i) = b(\alpha, N, N) = \alpha b(\alpha, N-1, N-1)$$

and we assume $b(\alpha, N-1, k) = 0$ if $k > N-1$. Finally, we have $\sum_{i=0}^{S_c-1-x} b(\alpha_1, 0, i) = 1$. Thus,

$$\Delta_S = \sum_{y=S+1-S_c}^{\infty} p_1(y) \sum_{x=0}^{S_c-1} p_2(x) \alpha_1 b(\alpha_1, y-S+S_c-1, S_c-1-x) + \sum_{i=0}^{S-S_c-1} p_1(i) p_2(S-i)$$

which is always positive. $\square$

Appendix 4: Proof of Lemma 2

Let $\Delta_{S_c} = \beta_1^c(S, S_c+1) - \beta_1^c(S, S_c)$. Also denote $p_1(y) = p(y; (\lambda_1 + \lambda_2)(L-T))$ and $p_2(x) = p(x; \lambda_1 T)$.

Then, we have

$$\begin{aligned} \Delta_{S_c} &= \sum_{y=S-S_c-1}^{\infty} p_1(y) \sum_{x=0}^{S_c} p_2(x) \sum_{i=0}^{S_c-x} b(\alpha_1, y-S+S_c+1, i) \\ &\quad - \sum_{y=S-S_c}^{\infty} p_1(y) \sum_{x=0}^{S_c-1} p_2(x) \sum_{i=0}^{S_c-1-x} b(\alpha_1, y-S+S_c, i) \\ &\quad + \sum_{i=0}^{S-S_c-2} \sum_{x=0}^{S-i-1} p_1(i) p_2(x) - \sum_{i=0}^{S-S_c-1} \sum_{x=0}^{S-i-1} p_1(i) p_2(x) \\ &= p_1(S-S_c-1) \sum_{x=0}^{S_c} p_2(x) \sum_{i=0}^{S_c-x} b(\alpha_1, 0, i) + \sum_{y=S-S_c}^{\infty} p_1(y) p_2(S_c) b(\alpha_1, y-S+S_c+1, 0) \\ &\quad + \sum_{y=S-S_c}^{\infty} p_1(y) \sum_{x=0}^{S_c-1} p_2(x) b(\alpha_1, y-S+S_c+1, S_c-x) \\ &\quad + \sum_{y=S-S_c}^{\infty} p_1(y) \sum_{x=0}^{S_c-1} p_2(x) \left[\sum_{i=0}^{S_c-1-x} b(\alpha_1, y-S+S_c+1, i) - \sum_{i=0}^{S_c-1-x} b(\alpha_1, y-S+S_c, i) \right] \\ &\quad - p_1(S-S_c-1) \sum_{x=0}^{S_c} p_2(x) \end{aligned}$$

Note that

$$\sum_{i=0}^{S_c-1-x} b(\alpha_1, y-S+S_c+1, i) - \sum_{i=0}^{S_c-1-x} b(\alpha_1, y-S+S_c, i) = -\alpha_1 b(\alpha_1, y-S+S_c, S_c-1-x)$$

and $\sum_{i=0}^{S_c-x} b(\alpha_1, 0, i) = 1$. Thus, we have

$$\begin{aligned} \Delta_{S_c} &= \sum_{y=S-S_c}^{\infty} p_1(y) p_2(S_c) b(\alpha_1, y-S+S_c+1, 0) \\ &\quad + \sum_{y=S-S_c}^{\infty} p_1(y) \sum_{x=0}^{S_c-1} p_2(x) [b(\alpha_1, y-S+S_c+1, S_c-x) - \alpha_1 b(\alpha_1, y-S+S_c, S_c-1-x)] \end{aligned}$$

Realizing that

$$b(\alpha, N, k) - \alpha b(\alpha, N-1, k-1) = \begin{cases} b(\alpha, N, k)(N-1)/N, & \text{if } k < N \\ 0, & \text{if } k \geq N \end{cases}$$

Δ_{S_c} is always positive. $\square$

References

- [1] Atkins, D. and Katircioglu, K. K. (1995) Managing inventory for multiple customers requiring different levels of service. Working Paper 94-MSC-015, University of British Columbia, Vancouver, BC.
- [2] Cohen, M. A., Kleindorfer, P. and Lee, H. L. (1989) Service constrained (s, S) inventory systems with priority demand classes and lost sales. *Management Science*, 34:482–499.
- [3] De Vericourt, F., Karaesmen, F. and Dallery, Y. (2002) Optimal stock allocation for a capacitated supply system. *Management Science*, 48:1486–1501.
- [4] Dekker, R., Hill, R. M., Kleijn, M. J. and Teunter, R. H. (2002) On the $(S - 1, S)$ lost sales inventory model with priority demand classes. *Naval Research Logistics*, 49:593–610.
- [5] Dekker, R., Kleijn, M. J. and Rooij, P. J. D. (1998) A spare parts stocking system based on equipment criticality. *International Journal of Production Economics*, 56–57:69–77.
- [6] Deshpande, V., Cohen, M. A. and Donohue, K. (2003) A threshold inventory rationing policy for service-differentiated demand classes. *Management Science*, 49:683–703.
- [7] Evans, R. V. (1968) Sales and restocking policies in a single item inventory system. *Management Science*, 14:463–472.
- [8] Frank, K. C., Zhang, R. Q. and Duenyas, I. (2003) Optimal policies for inventory systems with priority demand classes. *Operations Research*, 51:993–1002.
- [9] Ha, A. Y. (1997) Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science*, 43:1093–1103.
- [10] Ha, A. Y. (1997) Stock rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics*, 44:457–472.
- [11] Ha, A. Y. (2000) Stock rationing in an $M/E_k/1$ make-to-stock queue. *Management Science*, 46:77–87.
- [12] Hariharan, R. and Zipkin, P. (1995) Customer-order information, lead-times, and inventory. *Management Science*, 41:1599–1607.
- [13] Kaplan, A. (1969) Stock rationing. *Management Science*, 15:260–267, 1969.

- [14] Kleijn, M. J. and Dekker, R. (1998) An overview of inventory systems with several demand classes. In M. Grazia Speranza and P. Stahly (eds), “New Trends in Distribution Logistics”, Lecture Notes in Economics and Mathematical Systems 480, Springer Verlag, Berlin.
- [15] Kranenburg, A. A. and van Houtum, G. J. (2004) A multi-item spare parts inventory model with customer differentiation. Technical report, Beta Working Paper no. 110, Technische Universiteit Eindhoven, The Netherlands, ISBN 90-386-1938-3.
- [16] Melchior, P. M., Dekker, R. and Kleijn, M. J. (1998) Inventory rationing in an (S, Q) inventory model with lost sales and two demand classes. *Journal of Operational Research Society*, 51(1):111–122.
- [17] Moynadeh, K. and Aggarwal, P. K. (1997) An information-based multi-echelon inventory system with emergency orders. *Operations Research*, 45:694–701.
- [18] Nahmias, S. and Demmy, W. (1981) Operating characteristics of an inventory system with rationing. *Management Science*, 27:1236–1245.
- [19] Simpson, K. F. (1958) In-process inventories. *Operations Research*, 6:863–872, 1958.
- [20] Teunter, R. H. and Haneveld, W. K. K. (1996) Reserving spare parts for critical demand. Technical report, Graduate School/Research Institute System, Organizations and Management (SOM), University of Groningen.
- [21] Topkis, D. M. (1968) Optimal ordering and rationing policies in a non-stationary dynamic inventory model with n demand classes. *Management Science*, 15:160–176.
- [22] Veinott, A. F. J. (1965) Optimal policy in a dynamic, single product, non-stationary inventory model with several demand classes. *Operations Research*, 13:761–778.
- [23] Wang, Y., Cohen, M. A., and Zheng, Y. S. (2002) Differentiating customer service on the basis of delivery lead times. *IIE Transactions*, 34:979–989.