IIE Transactions

ISSN: 0740-817X (Print) 1545-8830 (Online) Journal homepage: http://www.tandfonline.com/loi/uiie20

Spare parts inventory management with demand lead times and rationing

Y. Levent Koçağa & Alper Şen

To cite this article: Y. Levent Koçağa & Alper Şen (2007) Spare parts inventory management with demand lead times and rationing, IIE Transactions, 39:9, 879-898, DOI: 10.1080/07408170601013646

To link to this article: http://dx.doi.org/10.1080/07408170601013646

Published online: 28 Jun 2007.

IIE Transactions (2007) **39**, 879–898

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/07408170601013646

Spare parts inventory management with demand lead times and rationing

Y. LEVENT KOÇAĞA[1] and ALPER ŞEN[2,*]

[1] *Information and Operations Management Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089-0809, USA*
E-mail: kocaga@usc.edu
[2] *Department of Industrial Engineering, Bilkent University, Bilkent, Ankara 06800, Turkey*
E-mail: alpersen@bilkent.edu.tr

Received August 2004 and accepted August 2006

We study an inventory system that consists of two demand classes. The orders in the first class need to be satisfied immediately, whereas the orders in the second class are to be filled in a given demand lead time. The two classes are also of different criticality. For this system, we propose a policy that rations the non-critical 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. The service level for the critical class is an approximation, whereas the service level for the non-critical class is exact. We then conduct a computational study to show that our approximation works reasonably, the benefits of rationing can be substantial, and the incorporation of demand lead time provides more value when the demand class with demand lead time is the critical class. The research is motivated by the spare parts service system of a major capital equipment manufacturer that faces two types of demand. For this company, the critical down orders need to be satisfied immediately, while the less critical maintenance orders can be satisfied after a fixed demand lead time. We 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.

Keywords: Rationing, demand lead time, inventory management, spare parts

1. Introduction

The primary motivation behind this research is our experience with a leading capital equipment manufacturer. This 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 manufactures are very expensive investments and are critical to the operations of its customers. It is very costly to have unused capacity at a customer's manufacturing facility caused by equipment failure. The company has an extensive spare parts network to provide spare parts and service to customers to repair equipment failures and perform scheduled maintenance operations. The network consists of more than 70 company-owned distribution centers and depots across the globe. The company also has agreements with its leading customers to manage their stock rooms. Three regional distribution centers in North America, Asia, and 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 those customers who do not have stock rooms operated by the company in the event of equipment failures ("down orders"). The regional distribution centers may also be used as a primary source for down orders for certain customers. The regional distribution centers also 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 regional distribution centers are the primary source to meet these demands, but local depots can also be used for certain customers. Even though the maintenance activities are scheduled and known in advance by the customers, the capital equipment manufacturer we study does not have access to these schedules. Since each location supports many customers and a large installation base, the capital equipment manufacturer perceives these orders as random.

*Corresponding author

Both types of customer orders (down and lead time) go through an order fulfillment engine that 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 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, both consumable and non-consumable (more than 50 000 active parts need to be managed) and varying service level requirements for different customers.

Providing an implementable and "good" solution for the whole spares network is a proven challenge; we, however, focus on an important issue where improvements can provide immediate and significant benefits. At present, the company uses a regular base stock inventory system with one service level for all types of demand and does not account for demand lead time differences. 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. Examples include a distribution center facing demand from retailers as well as directly from end customers; a spare part that is used in equipment of varying criticality; or an item that is sold to many customers of different criticality. The reader is referred to Kleijn and Dekker (2000) for a comprehensive study illustrating various examples in which multiple demand classes arise.

Given a system with multiple demand classes, the easiest policy would be to use different stockpiles for each demand class. Inventory for each class could be managed separately to meet a different service level requirement. While this policy is practical and very appealing, the drawback is that no advantage could be gained from risk pooling and more safety stock would be needed. On the other hand, one could simply use the same pool of inventory to satisfy demand from various customer classes without differentiation. In this case, the total stock needed would be determined by the highest service level requirement. The drawback here is that the highest service level is offered to all demand classes, leading to increased inventory costs.

Rationing, or the so-called critical-level policy, lies between these two extremes. Rationing has proven effective for handling different demand classes with different stock–out costs or service levels. We will explain rationing assuming that there are two demand classes but the extension to several demand classes is straightforward. A part of the stock is reserved for high-priority demand: this 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 unsatisfied demand is backordered, one also has to decide how to handle arriving replenishment orders. Obviously, if there is a backorder for a high-priority customer upon the arrival of a replenishment order, an arriving replenishment order would be used to satisfy this backorder. In addition, if there is a backorder for a low-priority customer when a replenishment order arrives and the inventory level is at or above the critical level, one should use this replenishment order to satisfy this backorder. However, in the case of a low-priority backorder and an inventory level below the critical level, one can either satisfy this backorder or increase the inventory level. 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, determining which one of these is optimal depends on the problem settings. For example, if the backorder penalty is non-linear in the backorder length, it may be better to clear a low-priority backorder even though the inventory level is below the critical level. Note that the service level for the low-priority class is not affected by the way replenishment orders are handled.

Except for very specific cases, a simple critical-level policy with a static critical level will not be optimal. 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, not satisfying a non-critical customer demand may not be optimal, especially if the probability of a critical demand arrival within this time is very small. Therefore, an optimal policy should take into account the remaining lead times of outstanding replenishment orders. However, there are two difficulties in employing a dynamic rationing policy. First, rationing problems are theoretically difficult. In fact, the exact expressions for the service level and the inventory on hand cannot be derived even for the seemingly simple static rationing policy with two demand classes with Poisson arrivals, deterministic lead time and backordering. Therefore, the existing literature and most of the ongoing research on rationing are limited to static policies. The only exception in the literature on backordering is Teunter and Haneveld (1996), which uses a heuristic under a very restrictive assumption. Even if theoretical results were readily available, employing such a dynamic rationing policy would be extremely difficult from a practical point of view. In fact, the fulfillment engine (a commercial software) that is used in the capital equipment manufacturer we study is not capable of promising orders based on the status of replenishment orders. Thus, we prefer to focus on a static rationing policy where the critical level does not change over time.

While the specific industrial application in this study requires a higher service level for the demand class that has no demand lead time, it is possible that other applications require a lower service level for this demand class. Consider,

		Critical class	
		$DLT = 0$	$DLT = T$
Non-critical class	$DLT = 0$		✓
	$DLT = T$	✓	

Fig. 1. The four possible cases.

for example, a multi-channel retailer that 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. However, no prior commitment is made to the customers in the demand class without a demand lead time, who ask for inventory upon their arrival into the store. Obviously, the service level requirement for online customers would be higher than customers purchasing through the store.

We therefore study a more general model where each demand class is identified by two characteristics, namely its demand lead time requirement and its service level requirement. A demand class is either critical or non-critical (i.e., its service level requirement is either more or less than the other class) and its Demand Lead Time (DLT) is either zero or T. The four possible cases are illustrated in Fig. 1.

When both DLTs are zero, the problem is the classical rationing problem for which we give an overview of the existing literature in Section 2. When both DLTs are T, the problem can again be reduced to the classical rationing problem in which the replenishment lead time is reduced by the common DLT of T (see Hariharan and Zipkin, 1995). The two cases of interest in this paper are represented by the check marks in Fig. 1. For these cases, without loss of generality, we assume that class 1 has a DLT of zero, and class 2 has a DLT of T. Our analysis is general for the two cases: (i) service level requirement for class 1 is higher than class 2; (ii) service level requirement for class 2 is higher than class 1.

We model the system as a single-location system facing Poisson demand in two classes with rates λ_1 and λ_2, respectively. The spare parts inventory is replenished according to a $(S-1, S)$ policy, S being the order-up-to level. For simplicity, we consider a deterministic replenishment lead time, L. 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. The objective is to find the optimum S and S_c such that the given service level requirements $\bar{\beta}_1$ and $\bar{\beta}_2$ are satisfied.

The remainder of the paper is organized as follows. In Section 2, we review the literature on 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 to find the optimal base stock and critical levels that satisfy service level requirements. In Section 4, we present the results of our simulation study; these indicate that our approximation for the service level of the critical class works quite well for high service levels. In addition, we present the results of the optimization study which determines the settings where the rationing is most useful. These settings are when the non-critical demands are dominant in the arrival mix, when the service level requirements are significantly different and when the DLT is present for the critical class. Also in Section 4, we present our results on a case study using 64 parts from the capital equipment manufacturer that we described earlier. We conclude the paper in Section 5.

2. Literature review

We will review the literature on inventory systems with a DLT before elaborating on the literature about rationing. We will first focus on the periodic-review models and then proceed to the continuous-review models.

The concept of a DLT was first introduced by Simpson (1958), using the term "service time" for base stock, multi-stage production systems. Hariharan and Zipkin (1995) then coined the term "DLT" to describe inventory-distribution systems where customers do not require immediate delivery thus allowing a fixed delay. The key observation in both papers is that the DLT works just as the opposite of the supply lead time, reducing the inventory held for achieving the required service level. This fact also applies to our system, but the existence of the two service classes complicates the model. Moinzadeh and Aggarwal (1997) considered a two-echelon system with two modes of inventory replenishment. In their model all orders are satisfied on a first-come first-served basis and the two order classes differ only in their transportation lead times. We, however, consider a system where orders are satisfied on a first-due first-serve basis. Wang *et al.* (2002) analyzed 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 since it involves two classes of service differentiated by a DLT. Therefore, we will explore their work in detail.

Wang *et al.* (2002) first studied a single-location system and derived expressions for the inventory level distribution and random customer delay. They made a crucial observation: the service level for customers with positive DLTs is

higher than for customers with a zero DLT as long as there is a positive probability that the replenishment order corresponding to a customer with a positive DLT arrives before its demand due date. After deriving the steady-state performance metrics for the single-location system, the model was extended to a two-echelon system. By following an approach similar to the well-known METRIC, the multi-echelon network was decomposed into single-location subsystems. Analysis of the two-echelon setting showed that the system with two service classes results in significant inventory cost savings.

The literature about rationing begins with Veinott (1965), 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 critical-level policy. Topkis (1968) proved the optimality of this policy both for the backordering and lost sales cases, and showed that the critical-levels generally decrease with the remaining time until the next ordering opportunity. Evans (1968) and Kaplan (1969) independently derived the same results for two demand classes. Nahmias and Demmy (1981) derived expressions for the expected backorder levels for a multi-period model with zero lead times and an (s, S) inventory policy when a static critical level is used. Other work in periodic inventory models with multiple demand classes include Cohen *et al.* (1989), Atkins and Katircioglu (1995), and Frank *et al.* (2003).

Nahmias and Demmy (1981) 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 constant lead time and a critical-level policy, under the important assumption that there is at most one outstanding order. Melchiors *et al.* (1998) analyzed the same model with lost sales.

Deshpande *et al.* (2003) 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, thus making the allocation of arriving orders a major issue. They defined a "threshold clearing mechanism" to overcome the difficulty of allocating arriving orders and they provided an efficient algorithm for computing the optimal policy parameters that are defined by (Q, r, K), K being the threshold level.

Dekker *et al.* (1998) discussed a case study on the inventory control of infrequently needed 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 was assumed to be Poisson while the replenishment lead time was assumed to be deterministic. The primary contribution of this paper is the derivation of service levels for both classes in the form of a 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. Conversely, the service level for non-critical demand class is exact, since it is not affected by the way incoming orders are handled.

A relevant stream of research introduced by Ha (1997a) considers the limited production capacity for replenishment orders and analyzes the system through make-to-stock queues with multiple demand classes. Ha's initial model has two demand classes, exponential supply lead times and backordering. Extensions include multiple demand classes (Vericourt *et al.*, 2002), lost sales (Ha, 1997b), lost sales and Erlang lead time distributions (Ha, 2000), and lost sales and general lead time distributions (Dekker *et al.*, 2002).

Our study differs from earlier research in that we simultaneously consider DLTs 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 (2004) considered a single-location multi-item spare parts model with multiple demand classes. They developed a solution procedure based on Lagrange relaxation and reported 10–20% savings on inventory investment using real data from a semiconductor equipment manufacturer, ASML. Teunter and Haneveld (1996) studied a critical-level policy for two demand classes where the critical level depends on the remaining time until the next stock replenishment. The "remaining time policy" is characterized by a set of critical stocking times $(L_1, L_2, ...)$; if the remaining time until the next replenishment is between zero and 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 parts inventory system that faces two classes of demand arrivals. Class 1 demands are due immediately, whereas class 2 demands allow a deterministic DLT of T. 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. The service level requirement for class j is $\bar{\beta}_j, j = 1, 2$. In our model, we will use as our service level measure the type I service level, i.e., the probability of no stock-out. We note that because of the Poisson Arrivals See Time Averages property, this is also the type II service level, i.e., the

fill rate. As discussed in Section 1, we consider both cases: (i) $\bar{\beta}_1 > \bar{\beta}_2$; and (ii) $\bar{\beta}_2 > \bar{\beta}_1$.

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; and at its due date (T time units after its arrival), it is satisfied only if the on-hand inventory is above a critical level, S_c, otherwise 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 time 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 note that in this 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). Also, when class 2 is the critical class, the inventory is not reserved for the critical order as soon as it arrives. Whether the order will be satisfied at its due date (T units of time after its arrival) still depends on the availability of inventory at that time. When $\bar{\beta}_1 = \bar{\beta}_2$, no rationing is applied and we will later show that our model reduces to the deterministic replenishment lead time version of the model in Wang *et al.* (2002).

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 DLT longer than the replenishment lead times. This assumption is also valid for the capital equipment manufacturer that motivated this research.

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 we have to resort to approximations. In the next section, we will derive service level expressions for both classes. We will then use these expressions 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. For a given S and S_c, let $\beta_j^c(S, S_c)$ denote the service level for j, if class j is the critical class and let $\beta_j^n(S, S_c)$ denote the service level for the class j, if class j is the non-critical class. 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 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

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

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

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

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 H, the arrival of the $S - S_c$ units of total demand that has a negative impact on inventory, is in one of the two intervals $(t, t + L - T]$ or $(t + L - T, t + L]$ or after $t + L$:

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

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

$$F_1(S, S_c, y) = P\{H \le y\} = P\{D_1(t, t+y) + D_2(t, t+y) \ge S - S_c\}.$$

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

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

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

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

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

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 $f_2(S, S_c, y) = dF_2(S, S_c, y)/dy$ of the hitting time can be derived as

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

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:

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

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

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

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 Equations* (3) *and* (4) *is a lower bound for the actual critical service level, given that the priority clearing mechanism is employed, i.e., all incoming replenishment orders are allocated to the critical class until the on-hand inventory reaches* S_c.

Proof. Let $I(a)$ denote the inventory level net of backorders for the non-critical class, $B(a)$ denote the total backorders at time a. Also let $R(a, b]$ denote the replenishments that are received in the interval $(a, b]$. Since all outstanding replenishments at t will arrive at time $t + L$, we have the following:

$$I(t) - B(t) + R(t, t+H] + R(t+H, t+L] \ge S,$$

or

$$I(t) + R(t, t+H] \ge S - R(t+H, t+L] + B(t). \quad (5)$$

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,

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

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:

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

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

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

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:

$$I(t+L) \ge I(t+H) + R(t+H, t+L] - D_j(t+H, t+L] - x,$$

or

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

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

$$I(t+L) \ge 1. \quad \blacksquare$$

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 were developed in previous research. Note that we extend the single-echelon model studied in Wang *et al.* (2002) 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.* (2002), we will see that the the service levels for the critical and non-critical demand classes can both be expressed as

$$\begin{aligned}\beta_j^n(S,0) &= \beta_j^c(S,0)\\ &= P\{D_1(t,t+L] + D_2(t,t+L-T] \le S-1\}\\ &= \sum_{i=0}^{S-1} p(i;\lambda_1 L + \lambda_2(L-T)). \end{aligned} \tag{7}$$

The derivations are given in Appendix 1.

Note that our model extends the rationing models given in Dekker *et al.* (1998) and Deshpande *et al.* (2003) by introducing a DLT for the non-critical demand class. Dekker *et al.* (1998) study an $(S-1, S)$ inventory-rationing model and derive 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

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

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

which are same as the expressions given in Dekker *et al.* (1998) (with a slight change in notation).

In Deshpande *et al.* (2003), the authors consider a (Q, r) inventory policy in which non-critical demand is 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 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.* (2003) 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.

We now show some structural properties of the approximation. We first note that the following lemma is immediately clear from Equations (2) and (4). The lemma simply states that the service level for the critical class is higher than or equal to that of the non-critical class if the critical level is positive, and the service levels are equal otherwise.

Lemma 1. *$\beta_j^c(S,S_c) = \beta_k^n(S,S_c)$ if $S_c = 0$ and $\beta_j^c(S,S_c) \ge \beta_k^n(S,S_c)$ for all $S_c \ge 1$ for $j \ne k$.*

We next provide two lemmas that state 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 $\beta_j^c(S,S_c)$ given in Equation (4) is increasing in S.*

Proof. See Appendix 3. ■

Lemma 3. *The approximation for the critical service level $\beta_j^c(S,S_c)$ given in Equation (4) is increasing in S_c, if class 1 is the critical class, or if class 2 is the critical class and $\lambda_1 \ge \lambda_2$.*

Proof. See Appendix 4. ■

Note that $\lambda_1 \ge \lambda_2$ is only a necessary condition for the monotonicity of $\beta_1^c(S,S_c)$ in S_c for the case where class 2 is the critical class. We have observed throughout numerical study that for most of the problems with $\lambda_1 < \lambda_2$, monotonicity still holds.

3.2. *Service level optimization*

We use the previously derived critical and non-critical service levels to solve the following optimization problem to minimize inventory investment:

$$\min_{S,S_c} \; S, \tag{10}$$

$$\text{subject to} \quad \beta_j^c(S,S_c) \ge \delta_j \bar{\beta}_j, \qquad j = 1,2, \tag{11}$$

$$\beta_j^n(S,S_c) \ge (1-\delta_j)\bar{\beta}_j, \quad j = 1,2, \tag{12}$$

$$S, S_c \ge 0, \tag{13}$$

where

$$\delta_j = \begin{cases} 1, & \text{if } \bar{\beta}_j = \max_k \bar{\beta}_k,\\ 0, & \text{otherwise.}\end{cases}$$

The objective function in Equation (10) is the base stock level. If $\delta_j = 1$, class j is the critical class and the first constraint, Equation (11), states that the approximated service level for the critical demand class is higher than the minimum required service level. Note that this also ensures that the actual service level is also higher than the minimum required service level due to Theorem 1. If $\delta_j = 0$, class j is not the critical class and constraint (11) is redundant for class j. If $\delta_j = 0$, the second constraint, Equation (12), ensures that the actual service level of the non-critical demand class is higher than the minimum required service level. If $\delta_j = 1$, constraint (12) is redundant for class j. The third constraint, Equation (13), ensures the non-negativity of the base stock and critical levels.

Note that our objective is to minimize the base stock level S, as opposed to minimizing the average inventory on hand. First observe that unlike the case in a standard continuous-review $(S-1, S)$ policy, the inventory position is not equal to S in this system with DLTs. 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 class 2 demands that are not yet due. Subtracting the average inventory on order, the expected inventory level is then equal to $S - \lambda_1 L - \lambda_2(L - T)$. When we assume that fill rates are reasonably high (i.e., when the backorders are relatively small), we can approximate the expected inventory on hand by the expected inventory level (for which no known exact expression exists, not even for the zero DLT case). Therefore, we choose to minimize S (since $\lambda_1 L + \lambda_2(L - T)$ is constant), as an approximation to the true objective of minimizing the average inventory on hand. To test the accuracy of this approximation, we solved 58 problems where the service level for the critical class and the average inventory on hand is derived through simulation. The results indicate that for all problems, the (S^*, S_c^*) pair that minimizes the base stock level also minimizes the average inventory on hand. The detailed results can be seen in Appendix 5. In Section 4.4, we also show in a case study with 64 parts that the average backorder levels are very low with various high fill rates, verifying that $S - \lambda_1 L - \lambda_2(L - T)$ is a good approximation for the expected inventory on hand.

For the optimization problem given in Equations (10)–(13), the feasible region for the (S, S_c) can be reduced. First we determine the minimum and maximum values that S can take. The minimum value of S is the minimum amount of inventory needed to ensure that the non-critical service level requirement is satisfied without rationing ($S_c = 0$), i.e., $S_{\min} := \arg\min\{x \geq 0 : \beta_j^n(x, 0) \geq (1 - \delta_j)\bar{\beta}_j, j = 1, 2\}$. The maximum value of S is the minimum amount of inventory needed to ensure that the critical service level requirement is satisfied without rationing, i.e., $S_{\max} = \arg\min\{x \geq 0 : \beta_j^c(x, 0) \geq \delta_j\bar{\beta}_j, j = 1, 2\}$. In other words, $S_{\max}$ is the solution of the simple round-up policy. Note also 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 the non-critical service level requirement, $S - S_c$ should be at least $S_{\min}$. Therefore, it is sufficient to consider (S, S_c) pairs where $S_{\max} \geq S \geq S_{\min}$ and $S_c \leq S - S_{\min}$. This reduction in the feasible region is possible regardless of whether we use the approximation or simulation to derive the service level for the critical class. If we are using the approximation, we have shown in Lemma 3 that the service level for the critical class $\beta_j^c(S, S_c)$ is increasing in S_c for a given S if class 1 is the critical class, or if class 2 is the critical class and $\lambda_1 \geq \lambda_2$. In these cases, for a given S, it is enough to check whether $S_c = S - S_{\min}$ satisfies the critical service level requirement. Thus, the feasible region can further be reduced to (S, S_c) pairs where $S_{\max} \geq S \geq S_{\min}$ and $S_c = S - S_{\min}$. In Section 4.2 we solve the optimization problem given in Equations (10)–(13) using the approximation for the critical service level. Since this approximation is only a lower bound, there is an opportunity to further reduce the base stock by using the exact critical service level (derived through simulation). Simulation optimization results are also reported in Section 4.2.

4. Numerical study

Our numerical study is composed of four parts. In Section 4.1, we test the performance of the approximation for the critical service level that was suggested in Section 3.1 and identify the cases where we can estimate the actual service level with reasonable accuracy. To do this, we use a simulation model coded in C and compare the simulated service level with the service level calculated through the approximation. Having confirmed that the approximation works well in most cases, in Section 4.2 we use the approximation in the optimization model to demonstrate the impact of various factors on base stock levels and critical levels. We comment on the impact of the DLT on rationing in Section 4.3. In Section 4.4 we demonstrate our results using a dataset from the capital equipment manufacturer. We consider both cases throughout our numerical study: (i) the demand class with zero DLT (class 1) is critical; and (ii) the demand class with positive DLT (class 2) is critical. The case study only demonstrates the former case.

4.1. *Simulation study*

In this section, we compare the performance of the approximation for the critical service level to the actual (simulated) service level. All tables in this section show 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 (4), and the percentage difference (calculated as the percentage difference between the simulated critical service level and the approximation for the critical service level (i.e., $100 \times$ (simulation approximation)/simulation)). The two cases are reported in all tables. For the case where class 1 is critical ($c = 1, n = 2$), λ_c, λ_n, β_c, β_n refer to λ_1, λ_2, β_1, β_2, respectively. For the case where class 2 is critical ($c = 2, n = 1$), λ_c, λ_n, β_c, β_n refer to λ_2, λ_1, β_2, β_1, respectively.

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%. Such high service levels are quite common in industry, especially for critical parts or critical demand classes. Table 1 shows the performance of the approximation when the critical service level is around 99% for 19 different instances. In columns 5–8, we study the case where class 1 is critical. In columns 9–12, we study the case where class 2 is critical. The supply lead time, L, is 0.5 and the DLT, 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, we have seen 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 (not reported in the table) is 0.0005, which shows that our simulation can describe the system accurately. Observe that the approximation works

Table 1. Performance of the approximation for a fixed service level of 99% ($L = 0.5$ and $T = 0.1$)

				$c = 1, n = 2$				$c = 2, n = 1$			
λ_c	λ_n	S	S_c	β_n *(exact)*	β_c *(sim)*	β_n *(approx)*	*Percentage difference (%)*	β_n *(exact)*	β_c *(sim)*	β_c *(approx)*	*Percentage difference (%)*
1	4	5	3	0.3796	0.9995	0.9976	0.19	0.3084	0.9993	0.9976	0.17
2	4	6	3	0.5184	0.9981	0.9927	0.54	0.4695	0.9977	0.9927	0.50
3	4	7	3	0.6248	0.9968	0.9892	0.76	0.6025	0.9966	0.9891	0.74
4	4	8	3	0.7064	0.9962	0.9877	0.85	0.7064	0.9963	0.9877	0.86
5	4	9	3	0.7693	0.9958	0.9876	0.82	0.7851	0.9964	0.9877	0.88
6	4	10	3	0.8180	0.9958	0.9884	0.74	0.8436	0.9969	0.9885	0.83
7	4	11	3	0.8560	0.9960	0.9896	0.64	0.8867	0.9973	0.9898	0.75
8	4	12	3	0.8857	0.9964	0.9909	0.55	0.9181	0.9979	0.9913	0.66
9	4	13	3	0.9090	0.9967	0.9922	0.45	0.9409	0.9983	0.9927	0.57
10	4	14	3	0.9274	0.9971	0.9934	0.37	0.9574	0.9987	0.9940	0.47
11	4	15	3	0.9420	0.9975	0.9945	0.30	0.9693	0.9990	0.9951	0.39
12	4	16	3	0.9536	0.9978	0.9954	0.24	0.9779	0.9992	0.9961	0.32
2	4	8	1	0.9828	0.9983	0.9963	0.20	0.9756	0.9973	0.9957	0.16
3	4	8	2	0.9057	0.9974	0.9928	0.46	0.8946	0.9969	0.9927	0.43
4	4	8	3	0.7064	0.9962	0.9877	0.85	0.7064	0.9963	0.9877	0.86
5	4	8	4	0.4142	0.9943	0.9802	1.42	0.4335	0.9954	0.9802	1.53
6	4	8	5	0.1626	0.9923	0.9697	2.28	0.1851	0.9948	0.9697	2.52
7	4	8	6	0.0372	0.9910	0.9554	3.59	0.0477	0.9947	0.9554	3.96
8	4	8	7	0.0037	0.9921	0.9368	5.57	0.0055	0.9956	0.9367	5.91

quite well when the critical service level is around 99%. The average percentage differences between approximation and simulation are 1.13 and 1.18% for the two cases. Note also that the approximation works better for higher service levels and in fact the best 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 are handled, which is the major shortcoming of the approximation, is less important.

In Table 2, we repeat the analysis above for a critical service level around 95% for ten different instances. Again, the supply lead time, L, is 0.5 and the DLT, T, is 0.1. The approximation still works well, although the performance is not as good as the 99% service level case. The average percentage differences between the approximation and simulation are 1.43% and 3.46% for the two cases. 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 98%.

4.1.2. *Accuracy of the approximation with varying system parameters*

In Table 3, we allow the critical service level to vary and we test the performance of the approximation by varying a

Table 2. Performance of the approximation for a fixed service level of 95% ($L = 0.5$ and $T = 0.1$)

				$c = 1, n = 2$				$c = 2, n = 1$			
λ_c	λ_n	S	S_c	β_n *(exact)*	β_c *(sim)*	β_n *(approx)*	*Percentage difference (%)*	β_n *(exact)*	β_c *(sim)*	β_c *(approx)*	*Percentage difference (%)*
4	1	5	2	0.5697	0.9380	0.9190	2.03	0.6496	0.9609	0.9208	4.17
5	1	6	2	0.6696	0.9481	0.9339	1.50	0.7576	0.9712	0.9368	3.54
6	1	7	2	0.7442	0.9573	0.9467	1.11	0.8318	0.9790	0.9505	2.91
7	1	8	2	0.8006	0.9652	0.9573	0.82	0.8829	0.9850	0.9617	2.36
8	1	9	2	0.8436	0.9718	0.9658	0.62	0.9182	0.9892	0.9706	1.88
9	1	10	2	0.8769	0.9772	0.9726	0.47	0.9427	0.9923	0.9776	1.48
5	1	7	1	0.9258	0.9761	0.9722	0.40	0.9580	0.9883	0.9785	0.99
6	1	7	2	0.7442	0.9573	0.9467	1.11	0.8318	0.9790	0.9505	2.91
7	1	7	3	0.4532	0.9321	0.9118	2.18	0.5803	0.9666	0.9130	5.54
8	1	7	4	0.1851	0.9040	0.8671	4.08	0.2854	0.9517	0.8673	8.86

Table 3. Performance of the approximation with varying system parameters

						c = 1, n = 2				*c = 2, n = 1*			
S	S_c	λ_c	λ_n	L	T	β_n *(exact)*	β_c *(sim)*	β_n *(approx)*	*Percentage difference (%)*	β_n *(exact)*	β_c *(sim)*	β_c *(approx)*	*Percentage difference (%)*
7	2	6	2	0.5	0.1	0.6678	0.9486	0.9225	2.75	0.7442	0.9678	0.9258	4.34
8	2	6	2	0.5	0.1	0.8156	0.9773	0.9655	1.21	0.8705	0.9871	0.9678	1.96
9	2	6	2	0.5	0.1	0.9091	0.9909	0.9861	0.48	0.9421	0.9954	0.9874	0.80
10	2	6	2	0.5	0.1	0.9599	0.9967	0.9949	0.18	0.9769	0.9985	0.9955	0.30
11	2	6	2	0.5	0.1	0.9840	0.9989	0.9983	0.06	0.9917	0.9995	0.9986	0.10
5	2	1	1	1	0.5	0.8088	0.9950	0.9860	0.90	0.8088	0.9994	0.9989	0.05
5	2	2	1	1	0.5	0.5438	0.9481	0.9008	4.99	0.6767	0.9953	0.9906	0.48
5	2	3	1	1	0.5	0.3208	0.8377	0.7378	11.93	0.5438	0.9835	0.9662	1.75
5	2	4	1	1	0.5	0.1736	0.6961	0.5438	21.88	0.4232	0.9609	0.9208	4.17
5	2	5	1	1	0.5	0.0884	0.5614	0.3668	34.66	0.3208	0.9257	0.8543	7.71
5	2	1	1	1	0.5	0.8088	0.9950	0.9860	0.90	0.8088	0.9994	0.9989	0.05
5	2	1	2	1	0.5	0.6767	0.9936	0.9686	2.52	0.5438	0.9985	0.9972	0.13
5	2	1	3	1	0.5	0.5438	0.9928	0.9484	4.47	0.3208	0.9973	0.9946	0.27
5	2	1	4	1	0.5	0.4232	0.9923	0.9274	6.54	0.1736	0.9962	0.9914	0.48
5	2	1	5	1	0.5	0.3208	0.9921	0.9072	8.56	0.0884	0.9954	0.9880	0.74
14	3	10	4	0.5	0.10	0.9274	0.9971	0.9934	0.37	0.9574	0.9987	0.9940	0.47
14	3	10	4	0.5	0.20	0.9486	0.9983	0.9953	0.30	0.9863	0.9998	0.9975	0.22
14	3	10	4	0.5	0.30	0.9651	0.9990	0.9973	0.17	0.9972	1.0000	0.9996	0.04
14	3	10	4	0.5	0.40	0.9775	0.9994	0.9986	0.08	0.9997	1.0000	1.0000	0.00
14	3	10	4	0.5	0.50	0.9863	0.9995	0.9993	0.02	1.0000	1.0000	1.0000	0.00

single parameter such as the base stock level, the arrival rate for the critical demand class, the arrival rate for the non-critical demand class and the DLT. As seen from the first part of the table, the critical and non-critical service levels both increase as the base stock level increases. We also note that the difference between the actual and approximated service levels 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 rate, respectively. As we increase both rates, we see that both the critical and non-critical service levels deteriorate. As we observed previously, the performance of the approximation deteriorates as we begin to see lower service levels. The difference between the simulated and approximated critical service levels is at unacceptable levels for service levels around 60%. However, these service levels are rarely observed in practice, especially for critical items or for critical demand classes. In the fourth part of Table 3, we study the impact of the DLT, T. As T increases, both the critical and non-critical service levels increase. Again, the difference behaves as expected, attaining its smallest value when the critical service level is the highest. Also as expected, the non-critical service level is quite sensitive to the DLT, while the critical service level is insensitive to the DLT.

The results in our simulation study show that, with 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 and the service levels are high. In all of our experiments, the service level obtained through approximation is lower than the actual service level for the critical demand class as proven in Theorem 1. Finally, we observe that the performance of the approximation improves as the service level for the critical demand class increases; this is in line with the high service level needs for critical demand classes.

4.2. *Optimization study*

In this section, we present the outputs 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. Tables 4 and 5 show our results for various input parameters. In both tables, the first column represents the input parameter being considered. We again study two cases: (i) class 1 is critical; and (ii) class 2 is critical. The first case is shown in columns 2–8; the second in columns 9–15. For the first case, the second column represents the required base stock level if no rationing is used (the DLTs are still recognized). This base stock level is determined by the critical service level requirement (although we recognize DLTs, the policy without rationing is still a round-up policy). The third and fourth columns show the base stock level and the critical level that are found through the optimization study using the approximation for the critical service level. The fifth column shows the percentage saving resulting from

Table 4. Optimal parameters: approximation vs simulation ($\lambda_c = 1$, $L = 0.5$, $T = 0.1$, $\bar{\beta}_n = 0.80$ and $\bar{\beta}_c = 0.99$)

	c = 1, n = 2							*c = 2, n = 1*						
λ_n	S_r^*	S	S_c	*Percentage saving (%)*	S^*	S_c^*	*Percentage saving (%)*	S_r^*	S	S_c	*Percentage saving (%)*	S^*	S_c^*	*Percentage saving (%)*
1	5	4	1	20.00	4	1	20.00	5	4	1	20.00	4	1	20.00
2	6	5	2	16.67	5	2	16.67	6	5	2	16.67	5	2	16.67
3	6	6	0	0.00	5	1	16.67	7	6	2	14.29	6	2	14.29
4	7	6	2	14.29	6	2	14.29	8	7	2	12.50	6	1	25.00
5	8	7	2	12.50	6	1	25.00	8	7	2	12.50	7	2	12.50
6	8	7	2	12.50	7	2	12.50	9	8	2	11.11	7	1	22.22
7	9	8	2	11.11	7	1	22.22	10	8	2	20.00	8	2	20.00
8	10	8	2	20.00	7	1	30.00	11	9	2	18.18	8	1	27.27
9	10	9	2	10.00	8	1	20.00	12	10	2	16.67	9	1	25.00
10	11	9	2	18.18	8	1	27.27	12	10	2	16.67	9	1	25.00

using a rationing policy that uses the approximation for the critical service level compared to the round-up policy (100× (column 2–column 3)/column 2). The sixth and seventh columns show the true optimal values of the base stock level and the critical level derived through simulation. The eighth column shows the percentage saving resulting from using a rationing policy that uses the simulation results for the critical service level compared to the round-up policy (100× (column 2–column 6)/column 2). Columns 9–15 are defined similarly for the second case.

In Table 4, λ_n varies between one and ten while λ_c is fixed at one. For the first case ($c = 1, n = 2$), we can reach the optimal solution for four instances using our approximation; in other instances there is only a single unit gap. The good performance of the optimization study that uses approximation is attributed to relatively slow arrival rates and small lead time demands. Rationing tends to create more savings when the arrival rate in the non-critical demand class is significantly higher than in the critical demand class, although there is no uniformity in this behavior. This is intuitive, because there are more opportunities to ration, when there are more non-critical arrivals in the arrival mix. In this case, a policy without rationing becomes more inefficient, since a large fraction of customers will be supported by a higher service level than necessary. On the other hand, a rationing policy will utilize the increased proportion of customers who tolerate a lower service level through its ability to differentiate service and save inventory. Similar results are observed in the second part of the table for the second case ($c = 2, n = 1$).

In Table 5, $\bar{\beta}_c$ varies between 90 and 99.5%, while $\bar{\beta}_n$ is fixed at 80%. Out of 16 instances, the optimization model that uses the approximation for the critical service level can obtain the true optimal solution in two instances. For the remaining 14 instances, we see that approximation on the average can capture a significant portion of the savings possible through rationing. Clearly, rationing becomes more effective as the critical service level requirement increases. The reason is similar to that of the results obtained in Table 4. When the service level requirements are significantly different, a round-up policy will be very ineffective as the non-critical class is provided with a service level much higher than necessary. Thus, rationing is more valuable for these cases. The fact that benefits of rationing are more pronounced with higher values of $\bar{\beta}_c/\bar{\beta}_n$ (and higher values of λ_n/λ_c) is in line with the results in Deshpande *et al.* (2003).

As seen from Tables 4 and 5, more savings can be obtained if simulated service levels are used for the critical

Table 5. Optimal parameters: approximation vs simulation ($\lambda_c = 5$, $\lambda_n = 10$, $L = 2$, $T = 0.5$ and $\bar{\beta}_n = 0.80$)

	c = 1, n = 2							*c = 2, n = 1*						
$\bar{\beta}_c$	S_r^*	S	S_c	*Percentage saving (%)*	S^*	S_c^*	*Percentage saving (%)*	S_r^*	S	S_c	*Percentage saving (%)*	S^*	S_c^*	*Percentage saving (%)*
0.900	33	32	2	3.03	31	1	6.06	35	35	0	0.00	34	1	2.86
0.925	33	33	0	0.00	31	1	6.06	36	35	2	2.78	34	1	5.56
0.950	34	34	0	0.00	31	1	8.82	37	36	3	2.70	36	3	2.70
0.970	36	35	5	2.78	32	2	11.11	39	36	3	7.69	35	2	10.26
0.980	37	35	5	5.41	32	2	13.51	40	37	4	7.50	35	2	12.50
0.985	37	36	6	2.70	32	2	13.51	40	37	4	7.50	35	2	12.50
0.990	38	36	6	5.26	33	3	13.16	41	38	5	7.32	38	5	7.32
0.995	40	37	7	7.50	33	3	17.50	43	39	6	9.30	36	3	16.28

service level to optimize the base stock and critical levels. However, our optimization model, which uses the approximation, can be solved significantly faster than the simulation optimization. On a 2 GHz Pentium 4 processor, the run time of the optimization that uses the approximation 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 time can increase to several hours depending on system parameters. Considering the small differences between the base stock levels obtained, we conclude that our optimization model based on the approximation indeed performs very well in capturing most of the savings due to rationing, avoiding the computational burden of the simulation optimization. In addition, the optimization model that uses the approximation can serve another important role: the base stock level that is obtained through the approximation can be used as an upper bound for the simulation optimization, reducing the simulation time considerably.

4.3. *Importance of the DLT*

In this section, we study the impact of DLT differentiation on the benefits of rationing. Particularly, we investigate how the benefits of rationing vary depending on the magnitude of the DLT and depending on whether the demand class with DLT (class 2) is critical or non-critical. For this purpose, we first study the case where the arrival rates are identical. In Table 6, both arrival rates are set at ten. The replenishment lead time L is set at 0.5 and the DLT T varies between zero and 0.5. We study two cases: (i) class 1 is critical; and (ii) class 2 is critical. In Table 6, column 5 shows the base stock level if there is no rationing. Columns 6 and 7 show the base stock and critical level derived through simulation for the case where class 1 is critical. Column 8 shows the savings in base stock level against the round-up policy. Similarly we have columns 9, 10 and 11 for the case where class 2 is critical.

For a given T, the total net demand $(\lambda_1 L + \lambda_2(L - T))$ during lead time is the same for both cases. However, the proportion of net non-critical demand to the total demand during lead time is higher when class 2 is the critical class. Therefore, rationing is more beneficial when the critical demand class has a DLT. Thus, if there is an opportunity to incorporate a DLT into one of the demand classes, the critical demand class should be chosen rather the non-critical demand class. As T increases, this proportion increases for the case when class 2 is the critical demand, and decreases for the case when class 1 is the critical demand. However, as T increases, the total demand during lead time is not constant (thus neither is the base stock level for the round-up policy). Thus, for both cases, there is no clear trend in the benefits of rationing as T increases.

In order to keep the total net demand during the lead time constant, we vary λ_2 along with T such that $\lambda_2(L - T) = 5$ in Table 7. Since the total net demand during the lead time is constant, the round-up policy has the same base stock for all T. Also for all values of T, the critical net demand during the lead time is exactly equal to the non-critical net demand during lead time. Even though the net demand in both classes are exactly equal, the non-critical class is more dominant in the arrival mix when class 1 is the critical class. As discussed in Section 4.2, more arrivals in the non-critical class provide more opportunities for rationing, thus increasing the benefits of rationing. When class 1 is the critical class, as T increases, the proportion of non-critical arrivals also increases and we see more benefits from rationing. On the other hand, when class 2 is the critical level, the benefits of rationing decreases as T increases.

In Table 8, we vary both λ_1 and λ_2 along with T such that the total net demand during the lead time stays constant at 10. However, in this case, as T increases, the proportion of class 1 demand during the lead time to the total net demand during the lead time increases. Therefore, more benefits are

Table 6. Impact of the DLT: $\lambda_1 = \lambda_2$

					$\bar{\beta}_1 = 0.99, \bar{\beta}_2 = 0.80$			$\bar{\beta}_1 = 0.80, \bar{\beta}_2 = 0.99$		
λ_1	λ_2	L	T	S_r^*	S^*	S_c^*	*Percentage saving (%)*	S^*	S_c^*	*Percentage saving (%)*
10	10	0.5	0.00	19	17	3	10.53	17	3	10.53
10	10	0.5	0.05	18	16	3	11.11	16	3	11.11
10	10	0.5	0.10	18	15	3	16.67	15	3	16.67
10	10	0.5	0.15	17	15	3	11.76	15	3	11.76
10	10	0.5	0.20	16	14	3	12.50	14	3	12.50
10	10	0.5	0.25	16	14	3	12.50	14	3	12.50
10	10	0.5	0.30	15	13	3	13.33	13	3	13.33
10	10	0.5	0.35	14	12	2	14.29	12	2	14.29
10	10	0.5	0.40	13	12	3	7.69	11	2	15.38
10	10	0.5	0.45	13	11	3	15.38	10	2	23.08
10	10	0.5	0.50	12	11	3	8.33	9	1	25.00

Table 7. Impact of the DLT: constant total net demand during lead time and $\lambda_1 L = \lambda_2(L - T)$

					$\bar{\beta}_1 = 0.99, \bar{\beta}_2 = 0.80$			$\bar{\beta}_1 = 0.80, \bar{\beta}_2 = 0.99$		
λ_1	λ_2	L	T	S_r^*	S^*	S_c^*	*Percentage saving (%)*	S^*	S_c^*	*Percentage saving (%)*
10	10.00	0.5	0.00	19	17	3	10.53	17	3	10.53
10	11.11	0.5	0.05	19	17	3	10.53	17	3	10.53
10	12.50	0.5	0.10	19	16	2	15.79	17	3	10.53
10	14.29	0.5	0.15	19	16	2	15.79	17	3	10.53
10	16.67	0.5	0.20	19	16	2	15.79	17	3	10.53
10	20.00	0.5	0.25	19	16	2	15.79	17	3	10.53
10	25.00	0.5	0.30	19	16	2	15.79	17	3	10.53
10	33.33	0.5	0.35	19	16	2	15.79	18	4	5.26
10	50.00	0.5	0.40	19	16	2	15.79	18	4	5.26
10	100.00	0.5	0.45	19	15	1	21.05	18	4	5.26

obtained from rationing when in fact class 2 is the critical class (although this effect is visible only for the last two instances). Also, for this case, as T increases, the proportion of critical demand during the lead time increases, as do the benefits of rationing.

Overall, the magnitude of net non-critical demand during the lead time relative to the total net demand during the lead time plays a major role in determining the benefits of rationing. A secondary role is played by the relative magnitude of the non-critical arrival rate. Since we would like to have net non-critical demand during the lead time to be more dominant in the total net demand during the lead time, DLT or the service time should be incorporated in the critical class rather than the non-critical class. It is also important to highlight the extreme case of $T = L$ (the last rows of Tables 6 and 8). In this case, one would carry zero inventory for the class 2 customers, if separate stocks were to be used. Even then, pooling class 2 customers with class 1 customers can help to reduce the total stock.

4.4. *Case study*

This section shows the significance of our results using a case study at the capital equipment manufacturer that was briefly described in Section 1. We have selected a depot in North America that serves a number of customers for both down demand and lead time demand.

Table 9 summarizes the characteristics of 64 parts selected for our study. In order to ensure the appropriateness of the $(S - 1, S)$ inventory policy and the validity of the Poisson demand assumption, we included rather expensive and infrequently required parts in our study. We used the demand history of a 12 month period in 2001 and 2002 and included 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 varies for different parts. On the average, 52.2% of a part's demand is from down orders (i.e., critical demand). In the same 12 month period, these 64 parts had a sales volume of $41 200 000 (in cost).

Table 8. Impact of the DLT: constant total net demand during lead time and $\lambda_1 = \lambda_2$

					$\bar{\beta}_1 = 0.99, \bar{\beta}_2 = 0.80$			$\bar{\beta}_1 = 0.80, \bar{\beta}_2 = 0.99$		
λ_1	λ_2	L	T	S_r^*	S^*	S_c^*	*Percentage saving (%)*	S^*	S_c^*	*Percentage saving (%)*
10.00	10.00	0.5	0.00	19	17	3	10.53	17	3	10.53
10.53	10.53	0.5	0.05	19	17	3	10.53	17	3	10.53
11.11	11.11	0.5	0.10	19	17	3	10.53	17	3	10.53
11.76	11.76	0.5	0.15	19	17	3	10.53	17	3	10.53
12.50	12.50	0.5	0.20	19	17	3	10.53	17	3	10.53
13.33	13.33	0.5	0.25	19	17	3	10.53	17	3	10.53
14.29	14.29	0.5	0.30	19	17	3	10.53	17	3	10.53
15.38	15.38	0.5	0.35	19	17	3	10.53	17	3	10.53
16.67	16.67	0.5	0.40	19	17	3	10.53	17	3	10.53
18.18	18.18	0.5	0.45	19	17	3	10.53	16	2	15.79
20.00	20.00	0.5	0.50	19	17	3	10.53	15	1	21.05

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

$24 300 000 (59.1%) of this is by down orders; $16 000 000 (40.9%) by lead time orders.

The analysis is done in three steps. In the first step, we do not recognize the DLT 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. In this practice, the company places replenishment orders upstream for lead time orders when it ships the items to the customers, rather than at the demand arrival epoch. Therefore, currently, the base stock level is exactly equal to inventory on hand plus inventory on order minus backorders.

In the second step, we recognize the DLT for lead time orders, but do not use any rationing to provide differentiated service to the two types 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 Wang *et al.* (2002) 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 a differentiated service to the 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 the three different approaches. We see that recognizing the DLTs 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 DLTs saves 2.41% on total inventory (which is equal to the inventory investment, if we assume that the pipeline stocks are owned by the company); an 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 for the two types of demand. We also note that if the company was responsible for only on-hand inventory, the savings would be higher.

As the critical service level declines and approaches the non-critical service level, we see that savings due to the recognition of DLTs are still significant, while the impact of rationing is less pronounced.

Table 10. Impact of critical service level

	Crticial service level (%)			
	90	*95*	*97*	*99*
	80	*80*	*80*	*80*
No DLT—No rationing				
Base stock	11.449	12.294	12.930	14.050
Total inventory	11.517	12.324	12.946	14.054
On-hand inventory	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*
On-hand inventory	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*
On-hand inventory	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

It is also interesting that the average levels of backlogs are quite 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 ($41 200 000 for all parts). This justifies our approach to minimize base stock levels in our model.

We conclude that the recognition of the DLTs 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 important for this particular company) and as it is shown here, its performance is quite reasonable.

5. Conclusions

In this study, we consider a single-echelon spare parts distribution system with two demand classes. Orders in the first class must be satisfied immediately upon their arrival, whereas the orders in the second class can be satisfied after a fixed DLT. The two classes are also of different priority. We model the system as a single-echelon inventory model, where we propose a static rationing policy that would ration stock to the non-critical class.

We develop an approximation for the critical service level and prove that this approximation is essentially a lower bound for the critical service level. Then 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, especially for high critical service levels, which is in line with the needs of a critical demand class. In our optimization study, we show that rationing the non-critical orders results in significant savings in terms of base stock inventory. Savings are also demonstrated in a case study where we use real data from a capital equipment manufacturer.

We also present the situations where the rationing policy is more useful. The rationing policy is more effective when the net non-critical lead time demand dominates the net critical lead time demand, when the non-critical arrival rate dominates the critical arrival rate and when the critical service level requirement dominates the non-critical service level requirement. In addition, we also show that rationing provides more benefits when the DLT is incorporated to the critical class (in other words, advance demand information is more valuable if it can be obtained for customers in the critical class). Obviously even more savings would be possible by using the service levels that are obtained from simulation. However, our optimization algorithm is much faster than the simulation optimization, captures most of the savings due to rationing and hence, would be more effective, especially in systems with many parts (as for our capital equipment manufacturer).

To our knowledge, this study is the first to simultaneously consider rationing and DLT, both of which have independently proven to be cost effective for inventory systems. Our numerical results indicate that combining these two in fact 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, modeling more demand classes would be a useful extension. Other logical directions would be to consider several deterministic DLTs, stochastic supply or DLT, and inventory systems with a multi-echelon structure.

Acknowledgement

The authors would like to thank two anonymous referees and Abdullah Daşçı for their valuable feedback.

References

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.

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.

De Vericourt, F., Karaesmen, F. and Dallery, Y. (2002) Optimal stock allocation for a capacitated supply system. *Management Science*, **48**, 1486–1501.

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.

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.

Deshpande, V., Cohen, M.A. and Donohue, K. (2003) A threshold inventory rationing policy for service-differentiated demand classes. *Management Science*, **49**, 683–703.

Evans, R.V. (1968) Sales and restocking policies in a single item inventory system. *Management Science*, **14**, 463–472.

Frank, K.C., Zhang, R.Q. and Duenyas, I. (2003) Optimal policies for inventory systems with priority demand classes. *Operations Research*, **51**, 993–1002.

Ha, A.Y. (1997a) Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science*, **43**, 1093–1103.

Ha, A.Y. (1997b) Stock rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics*, **44**, 457–472.

Ha, A.Y. (2000) Stock rationing in an $M/E_k/1$ make-to-stock queue. *Management Science*, **46**, 77–87.

Hariharan, R. and Zipkin, P. (1995) Customer-order information, lead-times, and inventory. *Management Science*, **41**, 1599–1607.

Kaplan, A. (1969) Stock rationing. *Management Science*, **15**, 260–267, 1969.

Kleijn, M.J. and Dekker, R. (2000) An overview of inventory systems with several demand classes, in *New Trends in Distribution Logistics*, M.G. Speranza and P. Stahly, eds., Springer, Berlin, Germany.

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.

Melchiors, 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 the Operational Research Society*, **51(1)**, 111–122.

Moinzadeh, K. and Aggarwal, P.K. (1997) An information-based multi-echelon inventory system with emergency orders. *Operations Research*, **45**, 694–701.

Nahmias, S. and Demmy, W. (1981) Operating characteristics of an inventory system with rationing. *Management Science*, **27**, 1236–1245.

Simpson, K.F. (1958) In-process inventories. *Operations Research*, **6**, 863–872.

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.

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.

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.

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.

Appendices

Appendix 1

It is trivial to derive Equation (7) from Equations (2) and (4) when $S_c = 0$. In Wang *et al.* (2002), the service levels are given as Equations (26) and (27), which we reproduce below:

$$\beta_1 = P\{S - Q(0) + R(0) \geq 1\},$$
$$\beta_2 = P\{S - Q(0) + R(0) \geq 1\}$$

$$+ G_2(T)P\{S - Q(0) + R(0) = 0\},$$

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

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

Here, $G_1(.)$ and $G_2(.)$ 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 $T \le L$. 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 a mean value of zero. Also note that $G_2(T) = 0$ since $T \le L$. 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.* (2003), the service levels are given as Equations (13) and (14), which we reproduce below:

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

where

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

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

$$\beta_2 = 1 - \sum_{y=S}^{S}\sum_{x=y-S_c}^{\infty} p(x; \lambda L) = 1 - \sum_{x=S-S_c}^{\infty} p(x; \lambda L)$$
$$= \sum_{x=0}^{S-S_c-1} p(x; \lambda L), \tag{A1}$$

which is the 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

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

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

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

The first part of the sum can be derived using the same approach as in Deshpande *et al.* (2003). 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

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

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

Appendix 3

Proof of Lemma 2. Consider the expression for the critical service level that is given in Equation (4). First, using integration by parts:

$$\begin{aligned}&\int_0^{L-T} f_1(S, S_c, y)p(i : \lambda_j(L-y))\mathrm{d}y\\ &\quad = F_1(S, S_c, L-T)p(i : \lambda_j T)\\ &\qquad - \int_0^{L-T} F_1(S, S_c, y)\lambda_j\,[p(i : \lambda_j(L-y))\\ &\qquad - p(i-1 : \lambda_j(L-y))]\mathrm{d}y.\end{aligned}$$

Therefore, we have:

$$\begin{aligned}&\int_0^{L-T} f_1(S, S_c, y)\left(\sum_{i=0}^{S_c-1} p(i : \lambda_j(L-y))\right)\mathrm{d}y\\ &= F_1(S, S_c, L-T)\sum_{i=0}^{S_c-1} p(i : \lambda_j T) - \lambda_j\int_0^{L-T} F_1(S, S_c, y)\\ &\quad \times \sum_{i=0}^{S_c-1}[p(i : \lambda_j(L-y)) - p(i-1 : \lambda_j(L-y))]\mathrm{d}y\\ &= F_1(S, S_c, L-T)\sum_{i=0}^{S_c-1} p(i : \lambda_j T)\\ &\quad - \lambda_j\int_0^{L-T} F_1(S, S_c, y)p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y.\end{aligned}$$

Again using integration by parts:

$$\begin{aligned}&\int_{L-T}^{L} f_2(S, S_c, y)p(i : \lambda_j(L-y))\mathrm{d}y\\ &\quad = F_2(S, S_c, L)p(i : 0) - F_2(S, S_c, L-T)p(i : \lambda_j T)\\ &\qquad - \int_{L-T}^{L} F_2(S, S_c, y)\lambda_j[p(i : \lambda_j(L-y))\\ &\qquad - p(i-1 : \lambda_j(L-y))]\mathrm{d}y.\end{aligned}$$

Therefore, we have:

$$\begin{aligned}&\int_{L-T}^{L} f_2(S, S_c, y)\left(\sum_{i=0}^{S_c-1} p(i : \lambda_j(L-y))\right)\mathrm{d}y\\ &\quad = F_2(S, S_c, L) - F_2(S, S_c, L-T)\sum_{i=0}^{S_c-1} p(i : \lambda_j T)\\ &\qquad - \lambda_j\int_0^{L-T} F_1(S, S_c, y)\sum_{i=0}^{S_c-1}[p(i : \lambda_j(L-y))\\ &\qquad - p(i-1 : \lambda_j(L-y))]\mathrm{d}y\\ &\quad = F_2(S, S_c, L) - F_2(S, S_c, L-T)\sum_{i=0}^{S_c-1} p(i : \lambda_j T)\\ &\qquad - \lambda_j\int_0^{L-T} F_2(S, S_c, y)p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y.\end{aligned}$$

Note that $F_1(S, S_c, L-T) = F_2(S, S_c, L-T)$ and

$$\begin{aligned}F_2(S, S_c, L) &= P\{D_1(t, t+L] + D_2(t+T, t+L] \geq S - S_c\}\\ &= \sum_{i=S-S_c}^{\infty} p(i : \lambda_1 L + \lambda_2(L-T)).\end{aligned}$$

Thus, we have:

$$\begin{aligned}\beta_j^c(S, S_c) &= 1 - \lambda_j\int_0^{L-T} F_1(S, S_c, y)p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y\\ &\quad - \lambda_j\int_{L-T}^{L} F_2(S, S_c, y)p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y.\end{aligned} \tag{A2}$$

From Equation (A2), we have:

$$\begin{aligned}&\beta_j^c(S+1, S_c) - \beta_j^c(S, S_c)\\ &\quad = \lambda_j\int_0^{L-T}[F_1(S, S_c, y)\\ &\quad - F_1(S+1, S_c, y)]p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y\\ &\quad + \lambda_j\int_{L-T}^{L}[F_2(S, S_c, y)\\ &\quad - F_2(S+1, S_c, y)]p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y\\ &\quad = \lambda_j\int_0^{L-T} P\{D_1(t, t+y] + D_2(t, t+y]\\ &\quad = S - S_c\}p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y\\ &\quad + \lambda_j\int_{L-T}^{L} P\{D_1(t, t+y] + D_2(t+y-L+T, t+y]\\ &\quad = S - S_c\}p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y\\ &\quad = \lambda_j\int_0^{L-T} p(S - S_c : (\lambda_1+\lambda_2)y)p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y\\ &\quad + \lambda_j\int_{L-T}^{L} p(S - S_c : \lambda_1 y\\ &\quad + \lambda_2(L-T))p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y,\end{aligned}$$

which is always positive. ■

Appendix 4

Proof of Lemma 3. From Equation (A2), we have:

$$\begin{aligned}&\beta_j^c(S, S_c+1) - \beta_j^c(S, S_c)\\ &\quad = -\lambda_j\int_0^{L-T} F_1(S, S_c+1, y)p(S_c : \lambda_j(L-y))\mathrm{d}y\end{aligned}$$

$$+\lambda_j \int_0^{L-T} F_1(S, S_c, y)p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y$$
$$-\lambda_j \int_{L-T}^{L} F_2(S, S_c + 1, y)p(S_c : \lambda_j(L-y))\mathrm{d}y$$
$$+\lambda_j \int_{L-T}^{L} F_2(S, S_c, y)p(S_c - 1 : \lambda_j(L-y))\mathrm{d}y.$$

But, $F_1(S, S_c + 1, y) = F_1(S, S_c, y) + p(S - S_c - 1 : (\lambda_1 + \lambda_2)y)$ and $F_2(S, S_c + 1, y) = F_2(S, S_c, y) + p(S - S_c - 1 : \lambda_1 y + \lambda_2(L-T))$. Thus, we have:

$$\begin{aligned}
&\beta_j^c(S, S_c + 1) - \beta_j^c(S, S_c)\\
&= -\lambda_j \int_0^{L-T} F_1(S, S_c, y)[p(S_c : \lambda_j(L-y))\\
&\quad - p(S_c - 1 : \lambda_j(L-y))]\mathrm{d}y\\
&\quad - \lambda_j \int_{L-T}^{L} F_2(S, S_c, y)[p(S_c : \lambda_j(L-y))\\
&\quad - p(S_c - 1 : \lambda_j(L-y))]\mathrm{d}y\\
&\quad - \lambda_j \int_0^{L-T} p(S - S_c - 1 : (\lambda_1 + \lambda_2)y)p(S_c : \lambda_j(L-y))\mathrm{d}y\\
&\quad - \lambda_j \int_{L-T}^{L} p(S - S_c - 1 : \lambda_1 y\\
&\quad + \lambda_2(L-T))p(S_c : \lambda_j(L-y))\mathrm{d}y.
\end{aligned}$$

We solve the first term using integration by parts. Let

$$[p(S_c : \lambda_j(L-y)) - p(S_c - 1 : \lambda_j(L-y))]\mathrm{d}y = \mathrm{d}v \text{ and } F_1(S, S_c, y) = u.$$

Then,

$$\frac{1}{\lambda_j} p(S_c : \lambda_j(L-y)) = v \text{ and } f_1(S, S_c, y)\mathrm{d}y = \mathrm{d}u.$$

Hence,

$$\begin{aligned}
&\int_0^{L-T} F_1(S, S_c, y)[p(S_c : \lambda_j(L-y))\\
&\quad - p(S_c - 1 : \lambda_j(L-y))]\mathrm{d}y\\
&= \frac{1}{\lambda_j} F_1(S, S_c, L-T)p(S_c : \lambda_j T)\\
&\quad - \frac{1}{\lambda_j} \int_0^{L-T} f_1(S, S_c, y)p(S_c : \lambda_j(L-y))\mathrm{d}y.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\int_{L-T}^{L} F_2(S, S_c, y)[p(S_c : \lambda_j(L-y)) - p(S_c - 1 : \lambda_j(L-y))]\mathrm{d}y\\
&= -\frac{1}{\lambda_j} F_2(S, S_c, L-T)p(S_c : \lambda_j T)\\
&\quad - \frac{1}{\lambda_j} \int_{L-T}^{L} f_2(S, S_c, y)p(S_c : \lambda_j(L-y))\mathrm{d}y.
\end{aligned}$$

Recognizing $F_1(S, S_c, L-T) = F_2(S, S_c, L-T)$, we have:

$$\begin{aligned}
&\beta_j^c(S, S_c + 1) - \beta_j^c(S, S_c)\\
&= \int_0^{L-T} f_1(S, S_c, y)p(S_c : \lambda_j(L-y))\mathrm{d}y\\
&\quad + \int_{L-T}^{L} f_2(S, S_c, y)p(S_c : \lambda_j(L-y))\mathrm{d}y\\
&\quad - \lambda_j \int_0^{L-T} p(S - S_c - 1 : (\lambda_1 + \lambda_2)y)p(S_c : \lambda_j(L-y))\mathrm{d}y\\
&\quad - \lambda_j \int_{L-T}^{L} p(S - S_c - 1 : \lambda_1 y\\
&\quad + \lambda_2(L-T))p(S_c : \lambda_j(L-y))\mathrm{d}y.
\end{aligned}$$

Note that

$$f_1(S, S_c, y) = (\lambda_1 + \lambda_2)p(S - S_c - 1 : (\lambda_1 + \lambda_2)y)$$

and

$$f_2(S, S_c, y) = \lambda_1 p(S - S_c - 1 : \lambda_1 y + \lambda_2(L-T)).$$

Therefore,

$$\begin{aligned}
&\beta_j^c(S, S_c + 1) - \beta_j^c(S, S_c) = (\lambda_1 + \lambda_2 - \lambda_j)\\
&\quad \times \int_0^{L-T} p(S - S_c - 1 : (\lambda_1 + \lambda_2)y)p(S_c : \lambda_j(L-y))\mathrm{d}y\\
&\quad + (\lambda_1 - \lambda_j) \int_{L-T}^{L} p(S - S_c - 1 : \lambda_1 y\\
&\quad + \lambda_2(L-T))p(S_c : \lambda_j(L-y))\mathrm{d}y.
\end{aligned}$$

The first term above is always positive. The second term is zero if $\lambda_j = \lambda_1$ (i.e., class 1 is the critical class) and non-negative if $\lambda_j = \lambda_2$ (i.e., class 2 is the critical class) and $\lambda_1 \geq \lambda_2$. ■

Appendix 5

In Table A1, we study the impact of the choice of the objective function for the optimization model in Equations (10)–(13) as discussed in Section 3.2. We consider 58 instances and simulation is used to determine the service levels for the critical class during optimization. In the first part of the table, the objective is to minimize the average inventory on hand which is also derived through simulation. Optimal values of the base stock, critical level and average inventory on hand for this objective function are reported in columns 7–9, respectively. In the second part of the table, the objective is to minimize the base stock level, S. The optimal values of base stock level and the critical level for this objective function are reported in columns 10 and 11. As discussed in Section 3.2, $S - \lambda_1 L - \lambda_2(L-T)$ is an approximation for the average inventory on hand. This value is reported in column 12. We see that in all 58 problems, both objective functions result in the same base stock and critical values. In addition, the exact and approximated values of the average inventory on hand are very close. The results justify the use of base stock level as the objective function of the optimization model in Equations (10)–(13).

Table A1. Choice of objective function: average inventory on hand versus base stock

						Minimize AIOH			*Minimize S*		
λ_1	λ_2	L	T	$\bar{\beta}_1$	$\bar{\beta}_2$	S^*	S_c^*	*Avg. inv*	S^*	S_c^*	*Avg. inv.*
1	1	0.50	0.10	0.990	0.800	4	1	3.1091	4	1	3.10
1	2	0.50	0.10	0.990	0.800	5	2	3.7122	5	2	3.70
1	3	0.50	0.10	0.990	0.800	5	1	3.3303	5	1	3.30
1	4	0.50	0.10	0.990	0.800	6	2	3.9232	6	2	3.90
1	5	0.50	0.10	0.990	0.800	6	1	3.5575	6	1	3.50
1	6	0.50	0.10	0.990	0.800	7	2	4.1353	7	2	4.10
1	7	0.50	0.10	0.990	0.800	7	1	3.7737	7	1	3.70
1	8	0.50	0.10	0.990	0.800	7	1	3.4287	7	1	3.30
1	9	0.50	0.10	0.990	0.800	8	1	3.9895	8	1	3.90
1	10	0.50	0.10	0.990	0.800	8	1	3.6475	8	1	3.50
5	10	2.00	0.50	0.900	0.800	31	1	6.3856	31	1	6.00
5	10	2.00	0.50	0.925	0.800	31	1	6.3856	31	1	6.00
5	10	2.00	0.50	0.950	0.800	31	1	6.3856	31	1	6.00
5	10	2.00	0.50	0.970	0.800	32	2	7.4010	32	2	7.00
5	10	2.00	0.50	0.980	0.800	32	2	7.4010	32	2	7.00
5	10	2.00	0.50	0.985	0.800	32	2	7.4010	32	2	7.00
5	10	2.00	0.50	0.990	0.800	33	3	8.2940	33	3	8.00
5	10	2.00	0.50	0.995	0.800	33	3	8.3869	33	3	8.00
5	10	2.00	0.20	0.990	0.800	36	3	8.3500	36	3	8.00
5	10	2.00	0.40	0.990	0.800	34	3	8.3127	34	3	8.00
5	10	2.00	0.60	0.990	0.800	32	3	8.2758	32	3	8.00
5	10	2.00	0.80	0.990	0.800	29	2	7.3367	29	2	7.00
5	10	2.00	1.00	0.990	0.800	27	2	7.2936	27	2	7.00
5	10	2.00	1.20	0.990	0.800	25	2	7.2511	25	2	7.00
5	10	2.00	1.40	0.990	0.800	22	2	6.3180	22	2	6.00
5	10	2.00	1.60	0.990	0.800	20	2	6.2707	20	2	6.00
5	10	2.00	1.80	0.990	0.800	18	2	6.2178	18	2	6.00
5	10	2.00	2.00	0.990	0.800	17	3	7.0841	17	3	7.00
5	10	0.50	0.05	0.990	0.800	12	2	5.1642	12	2	5.00
5	10	0.50	0.10	0.990	0.800	12	2	5.6042	12	2	5.50
5	10	0.50	0.15	0.990	0.800	11	2	5.1304	11	2	5.00
5	10	0.50	0.20	0.990	0.800	10	2	4.6638	10	2	4.50
5	10	0.50	0.25	0.990	0.800	10	2	5.0984	10	2	5.00
5	10	0.50	0.30	0.990	0.800	9	2	4.6237	9	2	4.50
5	10	0.50	0.35	0.990	0.800	9	2	5.0340	9	2	5.00
5	10	0.50	0.40	0.990	0.800	8	2	4.5397	8	2	4.50
5	10	0.50	0.45	0.990	0.800	7	2	4.0999	7	2	4.00
5	10	0.50	0.50	0.990	0.800	7	2	4.5190	7	2	4.50
1	5	2.00	0.20	0.990	0.800	17	1	6.1257	17	1	6.00
1	5	2.00	0.40	0.990	0.800	16	1	6.1056	16	1	6.00
1	5	2.00	0.60	0.990	0.800	15	1	6.0867	15	1	6.00
1	5	2.00	0.80	0.990	0.800	14	1	6.0693	14	1	6.00
1	5	2.00	1.00	0.990	0.800	12	1	5.1020	12	1	5.00
1	5	2.00	1.20	0.990	0.800	11	1	5.0776	11	1	5.00
1	5	2.00	1.40	0.990	0.800	10	1	5.0556	10	1	5.00
1	5	2.00	1.60	0.990	0.800	9	1	5.0360	9	1	5.00
1	5	2.00	1.80	0.990	0.800	7	1	4.0497	7	1	4.00
1	5	2.00	2.00	0.990	0.800	6	1	4.0209	6	1	4.00
1	5	0.50	0.05	0.990	0.800	7	1	4.2819	7	1	4.25
1	5	0.50	0.10	0.990	0.800	7	1	4.5203	7	1	4.50
1	5	0.50	0.15	0.990	0.800	6	1	3.7865	6	1	3.75
1	5	0.50	0.20	0.990	0.800	6	1	4.0219	6	1	4.00
1	5	0.50	0.25	0.990	0.800	6	0	4.2529	6	0	4.25
1	5	0.50	0.30	0.990	0.800	5	1	3.5227	5	1	3.50
1	5	0.50	0.35	0.990	0.800	5	0	3.7525	5	0	3.75
1	5	0.50	0.40	0.990	0.800	4	1	3.0212	4	1	3.00
1	5	0.50	0.45	0.990	0.800	4	0	3.2518	4	0	3.25
1	5	0.50	0.50	0.990	0.800	3	1	2.5088	3	1	2.50

AIOH = average inventory on hand.

Biographies

Y. Levent Koçağa is currently a doctoral student in Operations Management in the Marshall School of Business at the University of Southern California. He holds B.S. and M.S. degrees in Industrial Engineering from Bilkent University, Ankara, Turkey. His current research interests are inventory planning and rationing policies for spare parts, revenue management and product line design with customization.

Alper Şen is an Assistant Professor of Industrial Engineering at Bilkent University, Ankara, Turkey. Prior to joining Bilkent University, he worked in high tech industries as a consultant in supply chain management. He holds B.S. and M.S. degrees in Industrial Engineering from Bilkent University and a Ph.D. degree in Operations Management from the Marshall School of Business, University of Southern California. His current research interests are inventory management for spare parts, revenue management and supply chain management. His work has appeared in *IIE Transactions*, *OR Letters* and *European Journal of Operational Research*.