Enabling customer satisfaction and stock reduction through service differentiation with response time guarantees

EI2016-13

Adriana F. Gabor
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, 3062 PA, Rotterdam. gabor@ese.eur.nl

Lars van Vianen
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, 3062 PA, Rotterdam.

Guangyuan Yang
Econometric and Tinbergen Institute, Erasmus School of Economics, Erasmus University Rotterdam, 3062 PA, Rotterdam. gyang@ese.eur.nl

Sven Axsäter
Department of Industrial Management and Logistics, Lund University, S-221 00, Lund. sven.axsater@iml.lth.se

In response to customer specific service time guarantee requirements, service providers can offer differentiated services. However, conventional customer differentiation models based on fill rate constraints do not take full advantage of the stock reduction that can be achieved by differentiating customers based on agreed response times. In this paper we focus on the $(S-1, S, K)$ model with two customer classes, in which low priority customers are served only if the inventory level is above K. We employ lattice paths combinatorics to derive the exact distribution of the response time (within leadtime) for the lower priority class and provide a simple and accurate approximation for the response time of the high priority class. We show that the stock levels chosen based on agreed response times can be significantly lower than the ones chosen based on fillrates. This indicates that response time guarantees are an efficient tool in negotiating after-sale contracts, as they improve customer satisfaction and reduce investment costs.

Key words: inventory planning, service differentiation, priority demand classes

1. Introduction

Due to high down time costs, operators of capital intensive equipment such as aircrafts, electronics and trucks, increasingly focus on the time needed to fix a failure and require response time guarantees. For example, Thales Netherlands, a supplier of naval radar and combat management systems, is required to provide a service level quantified as the maximum response time in case of a failure (van der Heijden et al. 2012). Unlike the fill rate, time based service levels enable providers to engage customers in customer-focused performance metrics, as they relate directly to down time costs (Cohen et al. 2006).

Even when they own the same product, different customers may require different time based service levels. (Cohen et al. 2006). For example, when a mainframe computer in a stock exchange fails, the financial impact will be more severe than when a mainframe in a library goes down. In such situations, service providers often categorize their customers in different priority classes, depending on the duration of the requested response time (Cohen et al. 2006) or the requested level of the fill rate (Arslan et al. 2007). Short response times and high fill rates correspond to high priority customers, while longer response times and lower fill rates correspond to the low priority customers.

The challenge to service providers is to find a way to comply with the service contracts for differentiated customers while having a minimal capital investment in service parts inventory. There are several ways to deal with inventory for differentiated customers. One way is to use separate pools of stocks for each demand class, which is less efficient than pooling stocks in one pool (Cohen et al. 2006). While pooling service parts without differentiation is more efficient, one has to deal with the free-rider problem: low priority customers may receive the same service level as high priority customers. In order to take advantage of the economies of scale of pooling while delivering differentiated services, researchers proposed to use critical level policies, that reserve a part of inventory for high priority customers and pool the rest of the resources (Veinott (1965), Nahmias and Demmy (1981), Dekker et al. (1998), Deshpande et al. (2003), Vicil and Jackson (2016), Arslan et al. (2007)). Most of the literature in the field focuses on minimizing expected on-hand inventory, while imposing a desired level on the fillrate. One drawback of critical level policies is that low priority customers may encounter long response times. One way to alleviate this problem would be to design contracts with response time guarantees in terms of probabilities, instead of, or additional to the fillrates. As response time guarantees are less strict than fillrate constraints, optimizing the stock levels based on them also leads to a decrease in stock levels.

In this paper we focus on the impact of response time constraints on the stock levels in a continuous-review $(S-1,S,K)$ inventory model, with two demand classes (Gold and Silver) and constant leadtime L. All the unsatisfied demands are backordered. Low priority (Silver) customers are served only when the on hand inventory is greater than K and replenishments are used to first clear Gold backorders, then to increase the reservation stock back to K, and finally to clear the Silver backorders. Such a policy is appropriate for differentiated spare parts services, characterized by low demand and items with high holding and shortage costs relative to the ordering costs (Dekker et al. 1998, Sherbrooke 1968, Alfredsson and Verrijdt 1999). Approximations for the fillrates in this system have been previously proposed in Deshpande et al. (2003) and Arslan et al. (2007), while recursive relations for approximating the steady state distributions of the number of customers of each type have been proposed by Vicil and Jackson (2016) and Fadıloğlu and Bulut (2010). To the best of our knowledge, the response time distributions for the two customer classes in this inventory model have not been previously characterized and the impact of response time constraints on stock levels has not been studied. Our contribution can be summarized as follows:

(i) We offer an exact derivation of the distribution of the response time (within leadtime) for the lower priority customers. Note that this distribution cannot be derived directly from the steady state distribution of the number of customers in the system, as it is the case in systems without priorities. The reason is that the waiting time of Silver customers depends not only on the total number of customers in the system seen upon arrival, but also on subsequent arrivals of Gold customers. To overcome this difficulty, we use elementary lattice paths counting, a technique that has often been used in the field of queuing theory by Champernowne (1956), Takács (1967), Böhm (2010). The advantage of this technique is that it leads to expressions that link naturally to the evolution of the system, as compared to the more analytical technique of Laplace Transforms, that is more common in the study of queues with priorities.

(ii) We propose a simple approximation for the response time distribution for Gold customers, based on a serial system similar to the one proposed by Arslan et al. (2007). This approximation also leads to a simple alternative method for calculating the fillrates in an $(S-1,S,K)$ inventory model.

As indicated in Vicil and Jackson (2016), the best performing approximations for the fillrates are the ones proposed by Vicil and Jackson (2016) and Fadıloğlu and Bulut (2010). In Section 7, we show that our simple approximation method gives very close results to those previously known, which are based on more complex recursive relations.

(iii) Via numerical experiments, we show that by using response time constraints instead of fillrates, a considerable reduction in stock levels can be achieved. This indicates that response time constraints can be an efficient tool in negotiating after sale contracts, as they lead to both customer satisfaction and low costs.

The paper is organized as follows. In Section 2, we review the literature on customer differentiation policies. In Section 3, we present our model and revise basic properties of the $(S-1, S, K)$ model. In Section 4, we use basic lattice path combinatorics to derive an explicit expression of the response time constraint for Silver customers and discuss an approximations for the response time distribution of Gold customers in Section 5. Al agorithm for deciding stock levels based on response time constraints is described in Section 6. In Section 7, we validate our approximation method for high priority customers via extensive numerical experiments and we discuss the impact of incorporating response time constraints on the stock levels. Conclusions and further reasearch directions are outlined in Section 8.

2. Literature review

Our paper relates at most to continuous review critical level inventory models with several demand classes and backordering. Critical level policies have been first proposed in Veinott (1965). Topkis (1968) analyzed this policy for a periodic system with zero leadtime and multiple demand classes, each with a different shortage cost. Each review period is divided into a finite number of subperiods, at the end of which the inventory manager allocates inventory to the demand realized so far. Topkis proves that within a review interval, there exist optimal, nonnegative, rationing levels for each demand class. Similar models have been analysed in Kaplan (1969) and Frank et al. (2003).

Ha (1997a,b) consider a make-to-stock single machine production system with several demand classes. For a Markovian model with Poisson demand and exponential production times, where the manager at the production facility has three possible actions (do not produce, produce one item to replenish or to satisfy a high priority backorder, and produce one item to fill a low priority backorder), he shows that a base stock policy for the production decision and a dynamic rationing policy for inventory are optimal.

Nahmias and Demmy (1981) are the first to propose approximations for the expected backorders and fillrates in an continous review (Q, R, K) inventory system with two demand classes and deterministic lead times. In this system, a (Q, R) policy is combined with a *priority clearing* policy, in which orders for the low priority customers are only satisfied when the inventory on hand is greater than K. Their approximation relies on the assumption that there is at most one outstanding order at any time, which implies that whenever a reorder quantity is received, the inventory position and inventory level are identical. This model is extended to multiple demand classes and compound Poisson demand in Moon and Kang (1998).

Dekker et al. (1998) give approximations for the fillrates in an $(S-1, S, K)$ model with deterministic lead times. They explore several ways of allocating the incoming replenishment items in case of stock out. They show that the allocation method has little influence on the fill rates, but impacts significantly the duration of the stock out for the lower priority class.

Deshpande et al. (2003) consider the continuous review (Q, R, K) inventory system with two customer classes discussed in Nahmias and Demmy (1981). They propose to approximate the system parameters with the optimal parameters in a *threshold clearing* mechanism that is easier to analyse. The authors show that this policy closely approximates the optimal priority clearing policy. Deshpande and Cohen (2005) extends the analysis of this model to multiple classes.

Arslan et al. (2007) show that the threshold clearing policy is equivalent to a pipeline allocation policy in which backorders for the higher priority class, orders to replenish the stock reserved for higher priority demand and backorders for lower priority demand are served according to a FCFS discipline. They show also that the inventory system using these policies can be analysed by mapping it to a serial inventory system and propose an efficient heuristic to find the policy parameters.

Vicil and Jackson (2016) analyse the $(S-1, S, K)$ inventory system with two demand classes, under the priority clearing policy. For exponential lead times, they propose an efficient recursive method for finding the steady state distribution of the on hand inventory and of the number of backorders of each class. They show that the same balance equations hold in case of general leadtimes, assuming that, for small h, the probability of a replenishment in $(t, t+h)$, is independent of the number of low priority backorders in the system. For constant lead times and assuming that the same independence condition holds, Fadıloğlu and Bulut (2010) propose a different recursive procedure to calculate the steady state probabilities, based on an embedded Markov chain. While the system we study is the same as the one in Vicil and Jackson (2016) and Fadıloğlu and Bulut (2010), the focus of our paper is on deriving the distribution of the response times for the two classes and studying the impact of using response times instead of fillrates on the stock levels.

Customer differentiation has also been studied in the context of queueing theory, however, mainly for single server systems or systems with a finite number of servers. The Laplace Stieltjes Transforms (LST) of the waiting times in a non-preemptive M/M/c queue with equal service rates have been derived by Davis (1966) and Kella and Yechiali (1985), while Kesten and Runnenburg (1957) and Miller Jr (1960) have derived the LST of the waiting time for the non-preemptive M/M/1 queue with different service rates. For $c = 1, 2$ the generating function of the number of low priority customers in a preemptive $M/M/c$ system has been recently derived in Wang et al. (2015). It is well known that the $(S-1, S)$ system can be modeled by an $M/G/\infty$ queue, in which arrivals coincide with order placements and service time coincides with the leadtime. Unlike in queuing theory, where priority concerns the order in which customers enter service, in an $(S-1, S, K)$ system, priority concerns the order in which arriving replenishments are given to customers. To the best of our knowledge, such a system has not been studied in the context of queuing theory.

3. Model description and preliminaries

We consider a service parts inventory system with two demand classes, i.e., high priority, or Gold customers, and low priority, or Silver customers. We will use indices G and Z to indicate the Gold and Silver priority classes. We assume that the arrival processes corresponding to the two customer classes are independent Poisson processes, with rates λ_G for Gold customers and λ_Z for Silver customers. The inventory system is controlled by a continuous review $(S-1, S, K)$ policy, that is, a base stock policy with a critical level K, that is characterized by the following rules: every time a demand from either a Gold or a Silver customer occurs, a replenishment order is placed with the producer; when the on-hand inventory is above the critical level K, both customer classes are served according to a first come, first served rule; when the inventory level is below K, all Silver orders are backlogged; when the on-hand inventory is depleted, all demand is backlogged. We assume $S \geq 1$. We define the *shortfall* for Gold customers as the amount needed to clear all Gold backorders and to restore the on-hand inventory to the critical level K. We assume that shortfalls for Gold customers have priority upon backorders of Silver customer, that is, pipeline items are used to satisfy first Gold backorders, then to increase the inventory level to K and finally to satisfy Silver back orders. The lead times are assumed to be non-negative and deterministic, denoted by L. Customers are differentiated by two service requirements: the response time τ_i, $i \in \{G, Z\}$, and the service level within the response time β_i, $i \in \{G, Z\}$, i.e., the proportion of customers satisfied within the response time. We assume that both response times are lower than the lead time, which is realistic for service providers.

We are interested in finding the minimal base stock level S for which there exists a reservation stock K, that ensures for each class i of customers, $i \in \{G, Z\}$, a response time τ_i with probability β_i. Note that minimal stock levels imply minimal expected stock on-hand and that the service level is guaranteed by the response time constraints. More precisely, our goal is to solve the following optimization problem

$$\begin{aligned} &\text{Min } S \\ &s.t. \ \Pr(R_G^K \leq \tau_G) \geq \beta_G \\ &\qquad \Pr(R_Z^K \leq \tau_Z) \geq \beta_Z \\ &\qquad S, K \in \mathbf{Z}_+. \end{aligned} \tag{1}$$

Below we present a list of parameters and notations that will be used throughout the paper.

S = base stock

L = lead time

K = critical level for Gold customers

λ_i = arrival rate of class i customers, $i \in \{G, Z\}$

λ = arrival rate of arbitrary customers

IL = inventory level

τ_i= reponse time for the class i customers, $i \in \{G, Z\}$

β_i= percentage of customers of priority class i, $i \in \{G, Z\}$ that will be served within τ_i

R_i = waiting time of a customer of priority class i, $i \in \{G, Z\}$ when $K = 0$

R_i^K = waiting time of a customer of priority class i, $i \in \{G, Z\}$ when $K > 0$

$\mathbf{po}(\cdot;\beta)$ - the probability mass function of a Poisson random variable with rate β

$\mathbf{Po}(\cdot;\beta)$ - the cumulative distribution function of a Poisson random variable with rate β

$\mathbf{bin}(\cdot;n,p)$ - the probability mass function of a Binomial distribution with parameters n and p

$\mathbf{Beta}(\cdot;\alpha,\gamma)$ - the cumulative distribution function of a Beta variable with parameters α and γ

$\mathbf{Erl}(\cdot;n,\lambda)$- the cumulative distribution function of an Erlang distribution with parameters n and λ

3.1. Preliminaries

The equivalent serial stage inventory model Our analysis relies on the equivalence between the $(S-1,S,K)$ inventory system and a serial stage inventory system (SSS). A similar serial system was used in Arslan et al. (2007) to analyse an $(S-1,S,K)$ inventory model where shortfalls for Gold and backorders for Silver are served in first come first served order.

The (SSS) inventory system divides the on hand inventory into 2 stockpiles (stages), each corresponding to one demand class. Let IL_i, $i = 1, 2$ be the inventory on hand at each stockpile and

$IL^{SSS} = IL_1 + IL_2$. Both stockpiles use a continuous review base stock policy, the first one with base stock level K and the second with level $S - K$. We call the stock at the first pile *the reservation stock*. The replenishment leadtime between the two stockpiles is assumed to be zero.

When a Gold customer arrives, if $IL_1 > 0$, he will be served from the first stockpile and a replenishment order is sent to the second stockpile. If $IL_2 > 0$, an item is sent to the first stock pile, thus restoring the inventory level. If $IL_2 = 0$, a *reservation* Gold backorder is registered at the first stockpile. If $IL_1 = IL_2 = 0$, a *real* Gold backorder is registered at the first stockpile. In both cases, a replenishment order is sent to the outside supplier. Observe that after this operation, $IL_1 = 0$ only when $IL_2 = 0$.

When a Silver customer arrives, he is directed to the second stockpile. If $IL_2 > 0$, he will be served, otherwise a Silver backorder at the second stockpile will be registered. In both cases, a replenishment order is placed with the outside supplier.

At the arrival of a replenishment item, two situations may occur. If there is a Gold backorder at the first stockpile, the item will fulfill a real Gold backorder, and if none is present a Gold reservation backorder. If no Gold backorder is registered at the first stockpile, the replenishment item will be given to a Silver backorder if one is registered, otherwise it will be used to replenish the stock at the second stockpile.

Figure 1 contains a schematic despription of the (SSS) system.

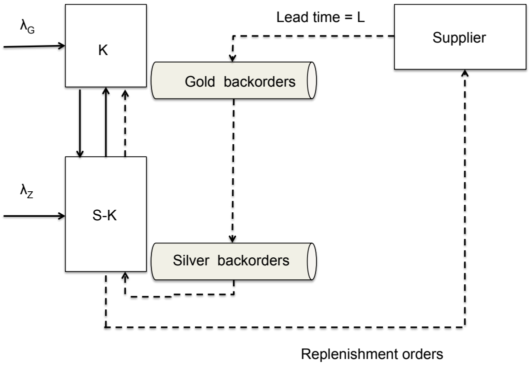

Figure 1 The equivalent serial system (SSS)

To show the equivalence between $(S-1, S, K)$ and (SSS) also holds in the case with pipeline priority, we follow the same approach as in Arslan et al. (2007). We assume that both systems start with full stock and argue that the on hand inventory and number of backorders of each class are the same every time a change in the system's state takes place: (i) when a demand occurs and (ii) when a replenishment arrives.

(i) *When a customer arrives.* If in the $(S-1,S,K)$ system, the inventory on hand is *larger* than K, it is decreased by one and a replenishment order is placed with the outside supplier. In the (SSS), $IL^{SSS} > K$ corresponds to $IL_2 > 0$ and $IL_1 = K$. Hence, when a demand from a Silver customer arrives, IL_2 is decreased by one. If the demand is placed by a Gold customer, IL_1 is first decreased by one, then immediately replenished, and IL_2 is decreased by one. In both cases, IL^{SSS} is decreased by one and a replenishment order is placed with the outside supplier.

If in the $(S-1,S,K)$ system, $0 < IL \leq K$, a demand from a Gold customer will be served, while a demand from a Silver customer will be backordered. The inventory level is decreased by one and a shortfall for Gold is registered. In the (SSS), this situation corresponds to $IL_2 = 0$, while $0 < IL_1 \leq K$. A demand from a Silver customer, that arives at the second stockpile, will be backordered, while a demand from a Gold customer will be served and a reservation Gold backorder will be registered at the first stockpile. In both cases, a replenishment order is placed with the outside supplier.

Finally, if there is no stock in the $(S-1,S,K)$ system, any incoming demand is backordered and a replenishment order is placed. In the (SSS), any Silver demand is backordered and any Gold demand triggers a real Gold backorder at the first stage. A replacement order is placed in both cases.

To sum up, when a demand arrives, on- hand inventory level in the two systems is the same, the number of Silver backorders in the $(S-1,S,K)$ system coincides with the number of Silver backorders at the second stage in the (SSS) system and the Gold shortfall in the $(S-1,S,K)$ system equals the Gold backorders (reservation and real) at the first stage.

(ii) *When a replenishment arrives* In the $(S-1,S,K)$ system, a replenishment will be first used to clear any shortfall for Gold (backorders and replenishment of reservation stock), and if the inventory on-hand is larger than K, it will be used for Silver backorders if any are present, or put in stock. It can be easily seen that in the SSS, a replenishment is used in a similar way.

As the $(S-1, S, K)$ and (SSS) systems are equivalent, we can use the second one to derive the waiting time distributions of the Gold and Silver customers. The response time distribution of Gold customers will be characterized based on the first stage of the (SSS) system, while the response time of Silver is based on the second stage. Observe that since at every Gold arrival a replenishment order is asked from the second stage, at the second stockpile, demand arrives at rate $\lambda_G + \lambda_Z$. Since when a replenishment item arrives, Gold backorders are fulfilled before Silver backorders, the second stage is equivalent to an $(S-K-1, S-K, 0)$ inventory model, with arrival rate $\lambda_G + \lambda_Z$.

Throughout the paper, we will be using the following Lemma to characterize the number of replenishment items in pipeline and the time till a certain replenishment will arrive in stock.

Lemma 1. *In an $(S-1, S, K)$ inventory system with Poisson arrivals and constant leadtime, the probability that a customer sees at arrival n items in pipeline is equal to $\boldsymbol{po}(n, \lambda L)$.*

Proof The proof relies on the well known equivalence between a $(S-1, S)$ (base-stock) inventory system with Poisson demands and an $M/G/\infty$ queue. The arrivals in the $M/G/\infty$ queue are the orders placed when customers arrive in the $(S-1, S)$ inventory system with Poisson demands. The service time is equivalent to the lead time. By Palm's theorem, the probability that a customer sees at arrival n items in pipeline is $\mathbf{po}(n, \lambda L)$. As priority only changes the order replenishment items are given to customers after they arrived in stock, the same holds in an $(S-1, S, K)$ system.

□

4. Response time distribution for Silver customers

In the previous section we have argued that the waiting time of Silver customers in an $(S-1, S, K)$ inventory model can be calculated based on the second stage of an (SSS) model, which behaves as a $(S-K-1, S-K, 0)$ model. In this model, both classes of customers are served as long as there is stock on hand, and Gold backorders have priority upon Silver backorders.

For the ease of the notation, consider a $(S-1,S,0)$ inventory system. Tag a Silver customer at his arrival, say time t. We take the arrival of the Silver customer as time reference. Assume that replenishment items in pipeline are numbered in increasing order of the residual lead time, that is, in the order they arrive to stock.

The Silver customer does not have to wait, if, at his arrival, there are items in stock, or equivalently, at most $S-1$ items in pipeline. Based on Lemma 1, we conclude that

$$P(R_Z=0)=\mathbf{Po}(S-1,\lambda L). \tag{2}$$

Let $N(u)$ be the number of replenishments that arrive in stock in $[t,t+u]$. Note that $N(u)$ is equal to the number of customer arrivals in $[t-L,t+u-L]$. Morover, let $N_G(u)$ represent the number of Gold arrivals in $[t,t+u]$. As for $u \in [0,L)$, the intervals $[t-L,t+u-L]$ and $[t,t+u]$ are disjoint, the variables $N(u)$ and $N_G(u)$ are independent. Thus, the process $(Y(u))_{u\in[0,L)}$, defined by $Y(u)=N(u)+N_G(u)$ can be seen as the restriction to $[0,L)$ of a Poisson process with rate $\lambda+\lambda_G$. Note that the process $(Y(u))_{u\in[0,L)}$ contains all the events (arrival of replenishment items and arrival of Gold customers) that impact the response time of Silver customers on $[0,L)$.

Assume that $N(L^-)=n$, where u^- denotes the time moment just before time u. That is, the tagged customer sees n items in pipeline upon arrival or, equivalently, $n-S$ customers waiting in front of him. In a system without priorities, the Silver customer would get the $n-S+1$-th replenishment item. However, in a system with priorities, the Silver customer may get the $n-S+j$-th replenishment with $j \geq 1$, due to Gold customers who are served before him based on the priority rule. Note that since the item ordered uppon the arrival of the tagged customer at t, will arrive in stock at $t+L$ and has label $n+1$, $j \leq S$ if the tagged customer is served in less than L time units. If the tagged customer gets the item ordered upon his arrival, $j=S+1$.

Let E_j be the event that the tagged customer gets the $n-S+j$-th replenishment item and define $p_{j,m,n}=P(E_j|N(L^-)=n,N_G(L^-)=m)$, for $n \geq S$ and $1 \leq j \leq S$.

By conditioning on E_j, $N(L^-)=n$ and $N_G(L^-)=m$, and taking into account that $N(u)$ and $N_G(u)$ are independent on $[0,L)$, we obtain that for $a \in (0,L)$,

$$P(0 < R_Z \leq a) = \sum_{n=S}^{\infty} \sum_{j=1}^{S} \sum_{m=j-1}^{\infty} P(0 < R_Z \leq a | E_j, N(L^-) = n, N_G(L^-) = m) p_{j,m,n} \mathbf{po}(n, \lambda L) \mathbf{po}(m, \lambda_G L). \tag{3}$$

Assume that event E_j takes place. The Silver customer gets the $n - S + j$-th replenishment item if $n - S + j$ replenishments and $j - 1$ arrivals of Gold customers take place in $[t, t + L)$. Hence, $m \geq j - 1$. Moreover, $R_Z = T_{n-S+2j-1}$, where $T_{n-S+2j-1}$ is the time when the $n - S + 2j - 1$ event takes place in the process $(Y(u))_{u \in [0,L)}$. Given that $Y(L^-) = n + m$, and Y is the restriction of a Poisson process on $[0, L)$, $T_{n-S+2j-1}$ is distributed as the $n - S + 2j - 1$-th order statistics of $n + m$ uniformly distributed random variables on $(0, L)$. In other words, $T_{n-S+2j-1}$ follows a $\mathbf{Beta}(\frac{\cdot}{L}, n - S + 2j - 1, m + S - 2j + 2)$ distribution.

The quantities $p_{j,n,m}$ are calculated in Lemma's 2 - 6 below. Let $t + A_k$ be the arrival time of k-th replenishment item and $N_G(A_k)$ denote the number of Gold customers that arrive in $[t, t + A_k]$.

Lemma 2. *Given that $N(L^-) = n$, event E_j occurs if and only if the following three conditions hold:*

(a) If $k \leq n - S + j - 1$, then $k \leq n - S + N_G(A_k)$

(b) $N_G(A_{n-S+j-1}) = j - 1$

(c) no Gold customer arrives in $[t + A_{n-S+j-1}, t + A_{n-S+j}]$.

Condition (a) states that at the arrival of k-th replenishment item, with $k \leq n - S + j - 1$, the number of replenishments could not cover the demand of the $n - S$ waiting customers and the number of Golds that arrived in the meantime. Condition (b) states that at the arrival of $n - S + j - 1$-th replenishment item, all the $n - S$ customers that were in queue at time t and the Gold customers that arived during $[t, t + A_{n-S+j}]$ have been served. The tagged customer will get item $n - S + j$ if and only if no Gold customer arrives between the arr (condition (c)). The proof of the Lemma 2 follows directly from these observations.

Definition 1. Let (a, b) and (p, q), with $a \leq p$, $b \leq q$ and $a, b, p, q \in \mathbf{Z}$, be two points in the euclidian plane. We call a path between (a, b) and (p, q) a *lattice path* if it starts in (a, b) and reaches (p,q) via unit length segments from left to right or upwards unit length segments. A lattice path with all the points on or below the line $y=x$ is called *subdiagonal.*

An example of a subdiagonal lattice path between (a,b) and (p,q) is given in Figure 2.

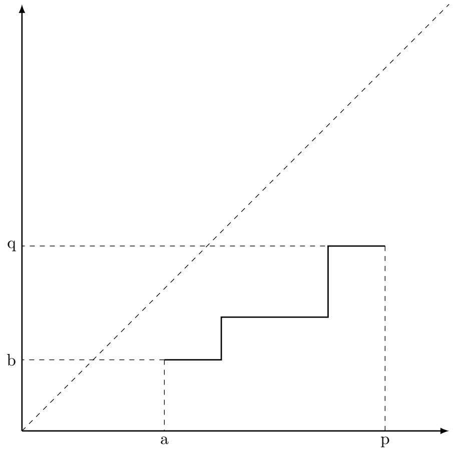

Figure 2 Subdigonal lattice path

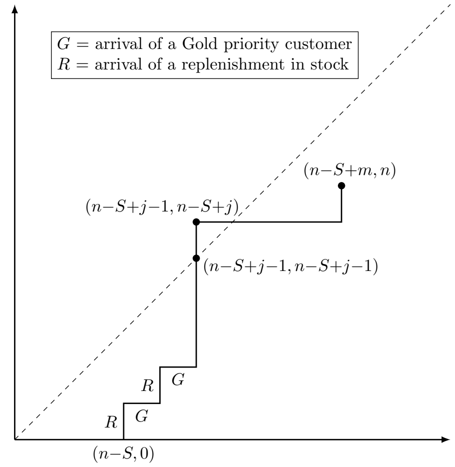
G = arrival of a Gold priority customer
R = arrival of a replenishment in stock
$(n-S+m,n)$
$(n-S+j-1,n-S+j)$
$(n-S+j-1,n-S+j-1)$
R
G
R
G
$(n-S,0)$

Figure 3 Example of a path in $\mathcal{L}_{m,n}$

Lemma 3. *(Brualdi (2004) Theorem 8.5.1) The number of lattice paths from (a,b) to (p,q), with $a \le p$ and $b \le q$ is equal to $\binom{p+q-a-b}{q-b}$.*

Lemma 4. *(Brualdi (2004) Theorem 8.5.3) Let p and q be integers with $p \ge q$. The number of subdiagonal lattice paths from $(0,0)$ to (p,q) is equal to $\frac{p-q+1}{p+1}\binom{p+q}{q}$.*

Assume that $N(L^-)=n$ and $N_G(L^-)=m$. Let the arrival of replenishment items to stock be labelled by R and the arrival of Gold items by G. Denote by $\mathcal{S}_{m,n}=\{(e_1,...,e_{m+n})|e_i \in \{R,G\}\}$. Each vector in $\mathcal{S}_{m,n}$ corresponds to a possible sequence of arrivals of replenishment items to stock and Gold customers during $[t,t+L]$. Let $\mathcal{L}_{m,n}$ be the set of lattice paths between $(n-S,0)$ and $(n-S+m,n)$ in the euclidian plane. Note that there exists a bijection between $\mathcal{S}_{m,n}$ and $\mathcal{L}_{m,n}$. To each element in $\mathcal{S}_{m,n}$ we associate a lattice path as follows: Start in $(n-S,0)$. Every time a replenishment arrives, draw an upwards vertical segment of unit length. Every time a Gold customer arrives, draw an horizontal segment of unit length from left to right. It is easy to see that each lattice path constructed in this way ends in $(n-S+m,n)$. Figure 3 shows an example of a path in $\mathcal{L}_{m,n}$.

LEMMA 5. *For $1 \le j \le S \le n$ and $j \le m+1$, the probability $p_{j,m,n}$ is given by*

$$p_{j,m,n} = \frac{1}{\binom{m+n}{m}} \frac{n-S+1}{n-S+j} \binom{n-S+2(j-1)}{j-1} \binom{m+S-2j+1}{S-j}.$$

Proof Recall that $p_{j,m,n} = P(E_j | N(L) = n, N_G(L) = m)$. According to Lemma 3, the number of lattice paths from $(n-S,0)$ to $(n-S+m,n)$ is equal to $\binom{m+n}{n}$, and since there is a bijection between $\mathcal{S}_{m,n}$ and $\mathcal{L}_{m,n}$, $|\mathcal{S}_{m,n}| = \binom{m+n}{n}$. Since the arrival of Gold customers and the arrival of replenishment items are independent on $[0,L)$, the probability that a sequence of events in $\mathcal{S}_{m,n}$ occurs is equal to $\frac{1}{\binom{m+n}{m}}$. To obtain $P(E_j|N(L) = n, N_G(L) = m)$ we thus only have to calculate the number of lattice paths in $\mathcal{L}_{m,n}$ that correspond to the event E_j.

Consider element $(x,y) \in \mathbf{Z} \times \mathbf{Z}$ on a lattice path from $(n-S,0)$ to $(n-S+m,n)$. Element (x,y) corresponds to the arrival of $x-(n-S)$ Gold customers and y replenishments, measured from the moment the tagged Silver customer arrived. Based on Lemma 2 a), $y \le n-S+j-1$ implies $y \le x$, or in other words, all the points (x,y) on the lattice path, with for $y \le n-S+j-1$, are subdiagonal. Condition (b) in Lemma 2 implies that the path touches the diagonal for the first time in $(n-S+j-1, n-S+j-1)$, while condition (c) implies that after the path crosses the diagonal, the next segment should be vertical. The lattice path then goes from point $(n-S+j-1, n-S+j)$ to point $(n, n-S+m)$.

Note further that by a mirroring argument, the number of subdiagonal lattice paths between $(n-S,0)$ and $(n-S+j-1, n-S+j-1)$ is equal to the number of subdiagonal lattice paths between $(0,0)$ and $(n-S+j-1, j-1)$. According to Lemma 4, this number equals $\frac{n-S+1}{n-S+j}\binom{n-S+2(j-1)}{j-1}$. Based on Lemma 3, the number of lattice paths from $(n-S+j-1, n-S+j)$ to $(n, n-S+m)$, for $m \ge S$, is equal to $\binom{m+S-2j+1}{S-j}$. Since each sequence of events corresponding to elements of $\mathcal{S}_{n,m}$ has the same probability of occuring, namely $\frac{1}{\binom{m+n}{m}}$, we obtain that:

$$p_{j,m,n} = \frac{1}{\binom{m+n}{m}} \frac{n-S+1}{n-S+k} \binom{n-S+2(j-1)}{j-1} \binom{m+S-2j+1}{S-j+1}.$$

□

Finally, $P(R_Z = L)$ can be obtained by conditioning again on the event that the Silver customer sees n items in pipeline upon arrival and he gets $n+1$-th replenishment item in pipeline:

$$P(R_Z = L) = \sum_{n=S}^{\infty} p_{n,S+1} \mathbf{po}(n, \lambda L), \tag{4}$$

with $p_{n,S+1}$ calculated in Lemma 6.

Lemma 6. *For $n \geq S$,*

$$p_{n,S+1} = \frac{n-S+1}{n+1} \boldsymbol{po}(S, \lambda_G L).$$

Proof To obtain $p_{n,S+1}$, note that the tagged Silver customer will get item $n+1$ if $m = S$ Gold customers have arrived in $[t, t+L]$. The set $\mathcal{L}_{S,n}$ has $\binom{n+S}{n}$ elements, hence the probability that an event in $\mathcal{S}_{S,n}$ takes place is $\frac{1}{\binom{n+S}{n}}$. Following the same reasoning as in Lemma 5, one can show that a sequence of events that leads to E_{S+1} can be represented as a subdiagonal lattice path from $(n-S, 0)$ to (n, n). The number of subdiagonal lattice paths from $(n-S, 0)$ to (n, n) is equal to the number of subdiagonal paths between $(0,0)$ and $(S, n-S)$, and by Lemma 4 is further equal to $\frac{n-S+1}{n+1}\binom{n+S}{n}$. Hence,

$$P(E_{S+1} | N(L) = n, N_G(L) = S) = \frac{n-S+1}{n+1}$$

and $p_{n,S+1} = \frac{n-S+1}{n+1} po(S; \lambda_G L)$.

□

To summarize, the distribution of R_Z, the response time of a Silver customer on $[0, L]$, is given by

Proposition 1. *In an $(S-1, S, 0)$ inventory model, the distribution of the response time R_Z of a Silver customer is given by*

$$P(R_Z = 0) = \boldsymbol{Po}(S-1, \lambda L),$$

$$P(R_Z \leq a) = \boldsymbol{Po}(S-1, \lambda L) + \sum_{n=S}^{\infty} \boldsymbol{po}(n, \lambda L) \sum_{j=1}^{S} \sum_{m=j-1}^{\infty} p_{j,m,n} \boldsymbol{Beta}(\frac{a}{L}; n-S+2j-1, m+S-2j+2), \tag{5}$$

for $a \in (0, L)$*, where*

$$p_{j,m,n} = \frac{n-S+1}{n-S+j}\binom{n-S+2(j-1)}{j-1}\frac{1}{\binom{m+n}{m}}\binom{m+S-2j+1}{S-j}\boldsymbol{po}(m, \lambda_G L)$$

and

$$P(R_Z = L) = \boldsymbol{po}(S, \lambda_G L)\sum_{n=S}^{\infty}\frac{n-S+1}{n+1}\boldsymbol{po}(n, \lambda L).$$

Despite the infinite sums and the combinatorial coefficients, the distribution of R_Z can be calculated very fast. In the numerical experiements discussed in Sections 7 and 7.3, we have evaluated $P(R_Z \leq a)$ for different values of a and $n = m = 500$ in a few milliseconds.

5. Response time distribution for Gold customers

In order to approximate the response time distribution for Gold customers, we return to the (SSS) serial stage inventory model. In this system, a Gold customer is served directly if the reservation stock at the first stage is not depleted. We distinguish two situations: when there is stock at the second stage and when there is not.

As the second stage works as a $(S-K-1, S-K, 0)$ system, the probability that a Gold customer sees upon arrival stock at the second stage is equal to the probability of seeing less than $S-K$ items in pipeline in the $(S-K-1, S-K, 0)$ system. Based on Lemma 1, this probability is equal to $\mathbf{Po}(S-K-1, \lambda L)$.

We focus now on the Gold customers who see upon arrival no stock at the second stage. They arrive according to a Poisson process with rate $\lambda_G \rho_2$, with $\rho_2 = 1 - \mathbf{Po}(S-K-1, \lambda L)$. To approximate the distribution of their response time, we approximate the first stage by an inventory model with base stock level K, where at the arrival of a Gold customer, a production order is placed with a supplier who has one exponential server with rate λ. The choice of the service rate is justified by the fact that the return process of replenishment is the arrival process of customers, shifted by a leadtime L. If a customer finds stock on hand upon his arrival, he will be immediately served, otherwise he will join the waitingline.

The production system at the supplier can thus be modelled by an $M/M/1$ queue with arrival rate $\lambda_G \rho_2$ and service rate λ. The stability of this $M/M/1$ queue is ensured by $\lambda_G < \lambda$. Note that the first K orders in the $M/M/1$ queue were placed by Gold customers who were actually served from the reservation stock .

Let R_G^K be the response time of a Gold customer in an $(S-1, S, K)$ system and let $\pi_k = \left(1 - \frac{\lambda_G \rho_2}{\lambda}\right)\left(\frac{\lambda_G \rho_2}{\lambda}\right)^k$, $k \in \mathbf{N}$ be the steady state probabilities in the $M/M/1$ queue with arrival rate $\lambda_G \rho_2$ and service rate λ.

For $K > 0$, a Gold customer will immediately be served if he either finds on hand inventory at the second stage or if he finds less than $K-1$ waiting orders in the $M/M/1$ queue. Thus,

$$\begin{aligned} P(R_G^K = 0) &\approx 1 - \rho_2 + \rho_2 \sum_{i=0}^{K-1} \pi_k. \\ &= 1 - \rho_2 + \rho_2(1 - \left(\frac{\lambda_G \rho_2}{\lambda})\right)^K) \\ &= 1 - \rho_2 \left(\frac{\lambda_G \rho_2}{\lambda}\right)^K. \end{aligned}$$

For $K = 0$, $P(R_G^K = 0) = \mathbf{Po}(S-1, \lambda L)$, which is the probability that in an $(S-1, S, 0)$ system there are items on stock.

A Gold customer will have to wait only if he sees $k \geq K$ waiting orders in the $M/M/1$ queue. In this case, as the first K orders are meant to restore the reservation stock at the first stage, the Gold customer will actually get the $k-K+1$-th replenishment item (where replenishment items are numbered in the order they arrive to stock). Thus, his response time is $\mathbf{Erl}(k-K+1, \lambda)$ distributed and

$$\begin{aligned} P(R_G^K < a) &\approx P(R_G^K = 0) + \rho_2 \sum_{k=K}^{\infty} \pi_k \mathbf{Erl}(k-K+1, \lambda) \\ &= 1 - \rho_2 + \rho_2 \sum_{i=0}^{K-1} \pi_k + \rho_2 \sum_{k=K}^{\infty} \pi_k (1 - \mathbf{Po}(k-K, \lambda a) \\ &= 1 - \rho_2 \sum_{k=K}^{\infty} \pi_k \mathbf{Po}(k-K, \lambda a) \\ &= 1 - \rho_2 \sum_{n=0}^{\infty} \pi_{n+K} \mathbf{Po}(n, \lambda a), \end{aligned}$$

where for the first equality we have used the well known identity $\mathbf{Erl}(a; i+1, \lambda) = 1 - \mathbf{Po}(i, \lambda a)$ and we assumed that $\sum_{i=0}^{K-1} \pi_k = 0$ for $K = 0$.

The quality of this approximation will be tested in Section 7.1.

6. Stock optimization algorithm

Next we describe a simple algorihm to find the optimal solution of the optimization problem (1). Let S^* and K^* be the minimal base stock and reservation level for which both the Gold and Silver response constraints are satisfied. Note that since the service level of a Silver customer is found based on an $(S-K-1, S-K, 0)$ system, a lower bound on $S^* - K^*$ is given by the the minimal base stock level for which $P(R_Z \leq \tau_Z) \geq \beta_Z$ in an $(S-1, S, 0)$ system. Denote by LB this lower bound. We further find the minimal positive U such that the Gold response constraint is satisfied in an $(LB+U-1, LB+U, U)$ system. Let $S^* = LB+U$. Clearly, S^* is the minimal base stock level for which both constraints are satisfied. However, as LB is a lower bound on $S^* - K^*$, the same service levels could be attained for reservation levels smaller than $S^* - LB$. We choose as K^* the minimal reservation level for which both response constraints are satisfied in the (S^*-1, S^*, K^*). Remark that $K^* \leq S^* - LB$. This algorithm will be used in Section 7.3 to study the impact of incorporating response time constraints on the stock levels.

7. Numerical experiments

In this section, we first validate the approximation of the response time distribution for Gold customers proposed in Section 5 and then we discuss the impact of response time constraints on the policy parameters.

7.1. Validation of the approximate response time distribution for Gold customers

The testbed of our experiments is similar to the one used in Arslan et al. (2007). The lead time is fixed to $L = 1/4$ year = 3 months.

To study the quality of the approximation proposed in Section 5, we evaluate the service level of Gold customers, $SL_G = P(R_G^K \leq \tau_G)$, for different combinations of S, K and τ_G as reported in Table 1 and Table 2. The upper part of Table 1 contains the results for $\lambda_G = \lambda_Z = 0.75$, while the lower part for $\lambda_G = \lambda_Z = 1.5$. The second column contains the values of the base stock levels $S \in \{4, 6, 8, 10, 12\}$ and the third column the values of the reservation stock $K \in \{0, 2, 4\}$. Columns 4-15 contain $SL_G = P(R_G^K \leq \tau_G)$ and the error made by the approximation, i.e., the difference between SL_G and the simulation, for $\tau_G \in \{0, 0.1, 0.25, 0.6, 0.75, 1\}$. For example, for $S = 4$, $K = 0$, the $SL_G = P(R_G^K \leq 0.1) = 0.413$ and the error is 0.79%. Table 2 contains similar information for the cases $\lambda_G = 0.75, \lambda_Z = 1.5$ (in the upper part) and $\lambda_G = 1.5, \lambda_Z = 0.75$ (in the lower part).

As the results show, in the cases we studied, the approximation is quite accurate, the maximum absolute error being 0.098. For equal arrival rates (see Table 1), the results are less accurate when the stock levels are low, i.e., $S \in \{4, 6\}$ for $\lambda_G = \lambda_Z = 0.75$ and $S \in \{4, 6, 8\}$ for $\lambda_G = \lambda_Z = 1.5$. However, the average error is around 1.35% in both cases, with a maximum error of 7.38% attained for for $S = 4$, $K = 0$, $\tau_G = 1$ and $SL_G = 0.83$. The lowest average error, of 0.9%, is obtained for the case $\lambda_G = 0.75, \lambda_Z = 1.5$, (see Table 2). In this case, the maximum error is 5%, obtained for $S = 6$, $K = 0$, $\tau_G = 0.75$ and $SL_G = 0.87$. The average error is the highest for the case $\lambda_G = 1.5, \lambda_Z = 0.75$, when it reaches a value of 3.1%. In this case, the maximum error is 9.85% and is obtained for $S = 4$, $K = 2$, $\tau_G = 1$ and $SL_G = 0.90$. In all cases, the highest error is obtained in cases where the service level is at most 0.90, which is unlikely to be desirable for the highest priority.

7.2. Comparison of fillrates with the fillrates obtained by other methods

As the numerical comparison in Vicil and Jackson (2016) shows, the best performing approximations for the fillrates for the two customer types in an $(S-1, S, K)$ system are the ones proposed by Vicil and Jackson (2016) and Fadıloğlu and Bulut (2010). The first one is based on solving recursively the balance equations of an approximate Cotinuous Markov Chain Model, (CTMC approach) while the second on solving the balance equations of an embedded Discrete Time Markov Chain model (DTMC approach). In both procedures, the transition probabilities are calculated via recursive procedures and one needs to truncate the systems of balance equations to find numerical solutions.

In Table 3 we compare our approximation with these two methods, by using the parameters used in Fadıloğlu and Bulut (2010). In all 18 experiments, $S = 4$ and $K \in \{1, 2\}$. The average number of arrivals during the leadtime $\lambda L \in \{1, 3, 6\}$. In our case, we chose $L = 3$ and $\lambda \in \{\frac{1}{3}, 1, 2\}$. Further, we chose the percentage of Gold customers such that $\frac{\lambda_G}{\lambda} \in \{0.25, 0.5, 0.75\}$. The information on

Table 1 Performance of the approximation method for the service level of Gold customers for equal Gold and Silver load

Case	Inputs		$\tau_G = 0$		$\tau_G = 0.1$		$\tau_G = 0.25$		$\tau_G = 0.6$		$\tau_G = 0.75$		$\tau_G = 1$	
	S	K	SL_G	$Error$[a]	SL_G[a]	$Err.$[b]	SL_G[a]	$Err.$[b]	SL_G[a]	$Err.$[b]	SL_G[a]	$Err.$[b]	SL_G[a]	$Err.$[b]
$\lambda_G = 0.75$	4	0	0.34	-0.08%	0.41	0.79%	0.51	2.17%	0.69	5.06%	0.75	6.04%	0.83	7.38%
$\lambda_Z = 0.75$	4	2	0.84	4.79%	0.86	4.95%	0.88	5.23%	0.93	5.56%	0.94	5.59%	0.96	5.51%
	4	4	0.97	3.43%	0.98	3.32%	0.98	3.19%	0.99	2.83%	0.99	2.67%	1.00	2.40%
	6	0	0.70	-0.07%	0.75	0.88%	0.81	2.09%	0.90	3.89%	0.93	4.22%	0.96	4.34%
	6	2	0.93	-0.33%	0.94	0.12%	0.95	0.66%	0.98	1.41%	0.98	1.56%	0.99	1.63%
	6	4	0.98	2.84%	0.99	2.75%	0.99	2.63%	0.99	2.26%	1.00	2.11%	1.00	1.84%
	8	0	0.91	0.01%	0.93	0.58%	0.95	1.18%	0.98	1.71%	0.99	1.69%	0.99	1.50%
	8	2	0.98	-1.49%	0.98	-1.14%	0.99	-0.73%	1.00	-0.18%	1.00	-0.05%	1.00	0.05%
	8	4	1.00	0.25%	1.00	0.27%	1.00	0.30%	1.00	0.30%	1.00	0.28%	1.00	0.25%
	10	0	0.98	0.00%	0.99	0.19%	0.99	0.37%	1.00	0.44%	1.00	0.41%	1.00	0.33%
	10	2	1.00	-0.41%	1.00	-0.30%	1.00	-0.19%	1.00	-0.06%	1.00	-0.03%	1.00	-0.01%
	10	4	1.00	-0.08%	1.00	-0.06%	1.00	-0.04%	1.00	-0.01%	1.00	-0.00%	1.00	0.00%
	12	0	1.00	0.01%	1.00	0.04%	1.00	0.07%	1.00	0.08%	1.00	0.07%	1.00	0.05%
	12	2	1.00	-0.06%	1.00	-0.04%	1.00	-0.02%	1.00	-0.00%	1.00	-0.00%	1.00	-0.00%
	12	4	1.00	-0.01%	1.00	-0.01%	1.00	-0.01%	1.00	-0.00%	1.00	-0.00%	1.00	-0.00%
$\lambda_G = 1.5$	4	0	0.02	0.02%	0.16	0.06%	0.34	0.68%	0.63	1.93%	0.71	2.43%	0.82	3.11%
$\lambda_Z = 1.5$	4	2	0.77	1.61%	0.80	1.65%	0.85	1.83%	0.92	1.97%	0.94	1.96%	0.96	1.88%
	4	4	0.95	1.43%	0.96	1.38%	0.97	1.33%	0.99	1.12%	0.99	1.01%	0.99	0.84%
	6	0	0.12	0.02%	0.26	0.34%	0.43	1.20%	0.71	3.14%	0.79	3.77%	0.88	4.41%
	6	2	0.78	1.87%	0.82	2.02%	0.86	2.27%	0.93	2.49%	0.95	2.47%	0.97	2.30%
	6	4	0.95	1.51%	0.96	1.46%	0.97	1.40%	0.99	1.18%	0.99	1.07%	1.00	0.88%
	8	0	0.32	0.04%	0.45	0.43%	0.60	1.26%	0.83	3.07%	0.88	3.48%	0.94	3.60%
	8	2	0.83	0.48%	0.86	0.95%	0.90	1.49%	0.96	2.04%	0.97	2.03%	0.99	1.85%
	8	4	0.96	1.69%	0.97	1.65%	0.98	1.56%	0.99	1.28%	0.99	1.14%	1.00	0.91%
	10	0	0.59	-0.03%	0.68	0.38%	0.78	1.09%	0.92	2.19%	0.95	2.24%	0.98	1.97%
	10	2	0.90	-2.53%	0.92	-1.63%	0.95	-0.64%	0.98	0.46%	0.99	0.60%	1.00	0.61%
	10	4	0.98	0.85%	0.98	0.90%	0.99	0.91%	1.00	0.78%	1.00	0.69%	1.00	0.53%
	12	0	0.80	-0.05%	0.85	0.45%	0.91	1.02%	0.98	1.35%	0.99	1.21%	1.00	0.87%
	12	2	0.95	-3.14%	0.96	-2.20%	0.98	-1.23%	0.99	-0.20%	1.00	-0.05%	1.00	0.05%
	12	4	0.99	-0.30%	0.99	-0.15%	1.00	-0.00%	1.00	0.12%	1.00	0.12%	1.00	0.09%

[a] $SL_G = P(R_G^K < \tau_G)$ service level of Gold customers obtained by simulation.
[b] $Err. = SL_G$ - service level of Gold customers obtained by simulation.

the input parameters is contained in columns 2, 3 and 4 in Table 3. In column 5 we report the exact value of the fillrate for Silver customers. Column 6 contains the approximate fillrate for Gold customers, calculated with the method proposed in Section 5 and column 7 contains the absolute error with respect to simulation. Finally, the last two columns report the absolute errors obtained by the CTMC and DTMC approaches.

As the procedure for calculating $P(R_Z^K = 0)$ is exact, we do not report the errors in the table. The approximation we propose for $P(R_G^K = 0)$ behaves slightly worse than the ones proposed by Vicil and Jackson (2016) and Fadıloğlu and Bulut (2010), however, the average error with respect

Table 2 Performance of the approximation method for the service level of Gold customers for different Gold and Silver loads

Case	Inputs		$\tau_G = 0$		$\tau_G = 0.1$		$\tau_G = 0.25$		$\tau_G = 0.6$		$\tau_G = 0.75$		$\tau_G = 1$	
	S	K	SL_G	$Error$[a]	SL_G[a]	$Err.$[b]	SL_G[a]	$Err.$[b]	SL_G[a]	$Err.$[b]	SL_G[a]	$Err.$[b]	SL_G[a]	$Err.$[b]
$\lambda_G = 0.75$	4	0	0.10	0.03%	0.24	0.72%	0.41	1.64%	0.68	3.27%	0.76	3.74%	0.85	4.21%
$\lambda_Z = 1.5$	4	2	0.90	1.03%	0.92	1.06%	0.94	1.07%	0.97	1.01%	0.98	0.98%	0.99	0.89%
	4	4	0.99	0.32%	0.99	0.31%	0.99	0.28%	1.00	0.22%	1.00	0.20%	1.00	0.16%
	6	0	0.33	0.07%	0.46	1.38%	0.60	2.90%	0.82	4.91%	0.87	5.15%	0.93	5.01%
	6	2	0.93	0.87%	0.94	1.01%	0.96	1.13%	0.98	1.17%	0.99	1.13%	0.99	0.99%
	6	4	0.99	0.41%	0.99	0.39%	1.00	0.35%	1.00	0.27%	1.00	0.24%	1.00	0.19%
	8	0	0.64	-0.02%	0.72	1.37%	0.81	2.76%	0.93	3.89%	0.96	3.73%	0.98	3.13%
	8	2	0.96	-0.72%	0.97	-0.37%	0.98	-0.00%	0.99	0.36%	1.00	0.39%	1.00	0.37%
	8	4	1.00	0.31%	1.00	0.30%	1.00	0.28%	1.00	0.20%	1.00	0.18%	1.00	0.13%
	10	0	0.86	0.01%	0.89	0.98%	0.93	1.79%	0.98	2.04%	0.99	1.81%	1.00	1.34%
	10	2	0.98	-1.05%	0.99	-0.72%	0.99	-0.40%	1.00	-0.05%	1.00	0.01%	1.00	0.04%
	10	4	1.00	-0.01%	1.00	0.01%	1.00	0.03%	1.00	0.03%	1.00	0.03%	1.00	0.02%
	12	0	0.96	0.06%	0.97	0.50%	0.98	0.80%	1.00	0.75%	1.00	0.63%	1.00	0.42%
	12	2	1.00	-0.43%	1.00	-0.30%	1.00	-0.16%	1.00	-0.03%	1.00	-0.01%	1.00	-0.00%
	12	4	1.00	-0.04%	1.00	-0.03%	1.00	-0.02%	1.00	-0.00%	1.00	0.00%	1.00	0.00%
$\lambda_G = 1.5$	4	0	0.10	0.03%	0.18	-0.73%	0.29	0.56%	0.51	3.58%	0.59	4.99%	0.70	7.38%
$\lambda_Z = 0.75$	4	2	0.63	6.25%	0.67	5.71%	0.72	7.03%	0.82	8.82%	0.85	9.30%	0.90	9.85%
	4	4	0.88	8.09%	0.90	7.60%	0.92	7.89%	0.95	7.77%	0.96	7.55%	0.98	7.05%
	6	0	0.34	0.08%	0.41	-0.05%	0.52	0.41%	0.71	2.72%	0.78	3.79%	0.86	5.26%
	6	2	0.71	4.13%	0.75	4.45%	0.79	5.58%	0.88	7.22%	0.91	7.58%	0.94	7.83%
	6	4	0.89	8.29%	0.91	7.86%	0.93	8.10%	0.96	7.89%	0.97	7.63%	0.98	7.04%
	8	0	0.64	0.01%	0.69	0.01%	0.77	0.39%	0.89	1.78%	0.92	2.21%	0.96	2.58%
	8	2	0.84	-2.91%	0.87	-1.91%	0.90	-0.64%	0.95	1.27%	0.97	1.71%	0.98	2.02%
	8	4	0.93	5.26%	0.94	5.26%	0.96	5.29%	0.98	4.99%	0.99	4.74%	0.99	4.20%
	10	0	0.86	0.03%	0.89	0.32%	0.92	0.71%	0.97	1.20%	0.98	1.24%	0.99	1.11%
	10	2	0.94	-4.25%	0.95	-3.27%	0.97	-2.12%	0.99	-0.57%	0.99	-0.23%	1.00	0.05%
	10	4	0.97	-0.17%	0.98	0.10%	0.99	0.37%	0.99	0.65%	1.00	0.66%	1.00	0.60%
	12	0	0.96	0.06%	0.97	0.28%	0.98	0.48%	0.99	0.56%	1.00	0.50%	1.00	0.37%
	12	2	0.98	-1.71%	0.99	-1.27%	0.99	-0.80%	1.00	-0.22%	1.00	-0.11%	1.00	-0.03%
	12	4	0.99	-0.68%	0.99	-0.50%	1.00	-0.29%	1.00	-0.07%	1.00	-0.02%	1.00	0.00%

[a] $SL_G = P(R_G^K < \tau_G)$ service level of Gold customers obtained by simulation.

[b] $Err. = SL_G$- service level of Gold customers obtained by simulation.

to simulation is 2% and the maximum error is 9.57% . The worst errors are obtained in cases 9, 14,15,17 and 18, that are characterized by large percentage of Gold customers (0.5% and 0.75%), very low fillrates for Silver customers (0.017, 0.062 and 0.199) and low fillrates for Gold customers (between 0.376 and 0.913). For the cases with a higher fillrate for Silver customers, that are likely to appear in practice, our approximation gives comparable results to the CTMC approach proposed by Vicil and Jackson (2016). The advantage of our approximation of the fillrates is that it is easier to implement and gives good results in situations relevant to practice.

Table 3 Comparison of fillrates of Gold customers with the fillrates obtained by the CTMC and the embedded DTMC approach for $S = 4$

Cases	K	λL	$\frac{\lambda_G}{\lambda}$	$P(R_Z^K = 0)$	$P(R_G^K = 0)$	Abs.Err.[a]	Abs.Err. CTMC[b]	Abs.Err.DTMC[c]
1	1	1	0.25	0.919	0.995	0.38%	0%	0%
2	1	1	0.5		0.990	0.67%	0%	0%
3	1	1	0.75		0.985	0.98%	0%	0%
4	1	3	0.25	0.423	0.911	0.58%	0.7%	0.1%
5	1	3	0.5		0.824	0.96%	1%	0%
6	1	3	0.75		0.737	1.37%	0.9%	0%
7	1	6	0.25	0.062	0.788	0.83%	2%	0.2%
8	1	6	0.5		0.580	1.99%	3.2%	0%
9	1	6	0.75		0.376	3.57%	2.8%	0.5%
10	2	1	0.25	0.735	0.998	0.04%	0.01%	0%
11	2	1	0.5		0.995	0.06%	0.05%	0%
12	2	1	0.75		0.989	0.07%	0.03%	0.1%
13	2	3	0.25	0.199	0.978	0.98%	0.52%	0%
14	2	3	0.5		0.913	4.11%	1.58%	0.1%
15	2	3	0.75		0.807	9.57%	2.19%	0%
16	2	6	0.25	0.017	0.948	0.71%	1.35%	0.1%
17	2	6	0.5		0.799	3.59%	4.21%	0%
18	2	6	0.75		0.562	9.53%	5.50%	0.7%

[a] *AbsErr.* = absolute error for the Gold fillrate between approximation and simulation
[b] *Abs.Err.* = absolute error for the Gold fillrate reported in Vicil and Jackson (2016)
[c] *Abs.Err.* = absolute error for the Gold fillrate reported in Fadıloğlu and Bulut (2010)

7.3. Impact of response time constraints on stock optimization

Next we analyse the impact of optimizing the stock levels based on reponse time constraints In our experiments, we use $\lambda_G = 0.75$, $\lambda_Z = 1.5$, $L = 3$ (months), $\beta_G = 0.99$ and $\beta_Z = 0.95$. For these input parameters, we use the algorithm described in Section 6 to solve the optimization problem (1). In Figure 4, we report the values for S^* and K^* for $\tau_G = 0$ and $\tau_G = 0.25$ months and for $\tau_Z \in [\tau_G, 3]$ months, where τ_Z is increased in steps of 0.05.

As we observe in the left graph in Figure 4, if the base stock level and the reservation stock are decided based on the fillrates (i.e., $\tau_G = \tau_Z = 0$), the optimal base stock is $S^* = 13$ and the optimal reservation stock is $K^* = 1$. However, if one agrees with the Silver customers on a response time of 0.28 months, S^* drops to 12, while $S^* = 11$ if the Silver customers are willing to wait for 0.68 months . This means a reduction of 8% and 15% in the base stock levels respectively. If one agrees with Gold customer on a response guarantee within 0.25 months, the stock level can be further reduced (see the right graph in Figure 4).

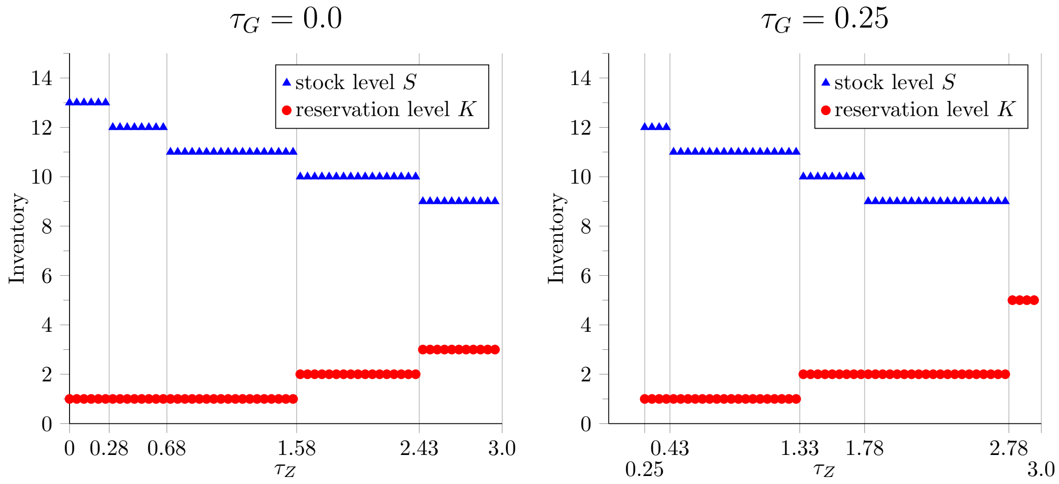

Figure 4 **Dependence of base stock levels on agreed response times**

8. Conclusions and further research

As after sale services become more and more competitive, it is important to develop models that incorporate customer oriented service measures into stock optimization problems. In this paper, we focused on the use of the response time as a measure of customer satisfaction and as a tool to reduce stock. In particular, we studied the impact of incorporating response time constraints on stock levels in an $(S-1, S, K)$ inventory model with two customer classes, Gold and Silver.

Our first result is an exact expression of the distribution of the response time (within leadtime) for Silver customers. The derivation is based on lattice path combinatorics, a technique that seems suitable to characterize other priority queues as well. The key impediment in extending our results beyond the leadtime is that on intervals $[0, t]$, with $t > L$, the return process of replenishments and the arrival of Gold customers are no longer independent. Although it is unlikely that Silver customers are willing to wait longer than the leadtime, from theoretical point of view, it would be interesting to have an analytical expression for the entire response time distribution for the low priority customers.

The second result is an easy approximation of the response time for Gold customers, based on an approximate two stage serial system. Via extensive experiments, we showed that this approximation performs very well, with and average error of 1.67% and a maximum error of 9.58%. The question of deriving the response time distribution for Gold customers remains an open question. A key assumption in our approximation is that waiting Gold customers are served at a constant rate λ, the arrival rate of replenishment items to stock. It would be interesting to study whether it is possible to circumvent this assumption without much loss in tractability.

Our numerical results show that incorporating response time constraints in optimizing an $(S-1, S, K)$ system can lead to significant decrease in stock levels. This indicates that response time constraints can be an important managerial tool in negotiating service contracts. By using the distribution of the response times derived in this paper, managers can offer clients a better indication of their waiting time than by using fillrates or expected waiting time. In our future research we will focus on extending this analysis to more complex inventory models.

Acknowledgements We thank Rommert Dekker (Erasmus School of Economics), Andrei Sleptchenko (Qatar University) and Jan Kees van Ommeren (University of Twente) for helpful discussions.

References

Alfredsson, P., J. Verrijdt. 1999. Modeling emergency supply flexibility in a two-echelon inventory system. *Management Science* **45**(10) 1416–1431.

Arslan, H., S.C. Graves, T.A. Roemer. 2007. A single-product inventory model for multiple demand classes. *Management Science* **53**(9) 1486–1500.

Böhm, Walter. 2010. Lattice path counting and the theory of queues. *Journal of Statistical Planning and Inference* **140**(8) 2168–2183.

Brualdi, RA. 2004. *Introductory Combinatorics*. 4th ed. Prentice Hall, NJ.

Champernowne, DG. 1956. An elementary method of solution of the queueing problem with a single server and constant parameters. *Journal of the Royal Statistical Society. Series B (Methodological)* 125–128.

Cohen, M. A., N. Agrawal, V. Agrawal. 2006. Winning in the aftermarket. *Harvard business review* **84**(5) 129.

Davis, R. H. 1966. Waiting-time distribution of a multi-server, priority queuing system. *Operations Research* **14**(1) 133–136.

Dekker, R., M.J. Kleijn, P.J. De Rooij. 1998. A spare parts stocking policy based on equipment criticality. *International Journal of Production Economics* **56** 69–77.

Deshpande, V., M.A. Cohen. 2005. A nested threshold inventory rationing policy for multiple demand classes in inventory systems with replenishment. Tech. rep., Working paper.

Deshpande, V., M.A. Cohen, K. Donohue. 2003. A threshold inventory rationing policy for service-differentiated demand classes. *Management Science* **49**(6) 683–703.

Fadıloğlu, M.M., Ö. Bulut. 2010. An embedded markov chain approach to stock rationing. *Operations Research Letters* **38**(6) 510–515.

Frank, K. C., R.Q. Zhang, I. Duenyas. 2003. Optimal policies for inventory systems with priority demand classes. *Operations Research* **51**(6) 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**(8) 1093–1103.

Ha, A.Y. 1997b. Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Research Logistics (NRL)* **44**(5) 457–472.

Kaplan, Alan. 1969. Stock rationing. *Management Science* **15**(5) 260–267.

Kella, O., U. Yechiali. 1985. Waiting times in the non-preemptive priority m/m/c queue. *Stochastic Models* **1**(2) 257–262.

Kesten, H, J Th Runnenburg. 1957. Priority in waiting line problems. ii. *Indagationes Mathematicae (Proceedings)*, vol. 60. Elsevier, 325–336.

Miller Jr, R. G. 1960. Priority queues. *The Annals of Mathematical Statistics* 86–103.

Moon, I, S Kang. 1998. Rationing policies for some inventory systems. *Journal of the Operational Research Society* 509–518.

Nahmias, S., W.S. Demmy. 1981. Operating characteristics of an inventory system with rationing. *Management Science* **27**(11) 1236–1245.

Sherbrooke, C.C. 1968. Metric: A multi-echelon technique for recoverable item control. *Operations Research* **16**(1) 122–141.

Takács, Lajos M. 1967. *Combinatorial methods in the theory of stochastic processes*, vol. 126. Wiley New York.

Topkis, D. M. 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science* **15**(3) 160–176.

van der Heijden, M.C., E.M. Alvarez, J.M.J. Schutten. 2012. Inventory reduction in spare part networks by selective throughput time reduction. *International Journal of Production Economics* **143**(2) 509–517.

Veinott, A.F. 1965. Optimal policy in a dynamic, single product, nonstationary inventory model with several demand classes. *Operations Research* **13**(5) 761–778.

Vicil, Oguzhan, Peter Jackson. 2016. Computationally efficient optimization of stock pooling and allocation levels for two demand classes under general lead time distributions. *IIE Transactions* (forthcoming).

Wang, Jianfu, Opher Baron, Alan Scheller-Wolf. 2015. M/m/c queue with two priority classes. *Operations Research* **63**(3) 733–749.