Comparison of two methods for customer differentiation

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

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 time guarantee requirements, service providers can offer differentiated services. However, conventional customer differentiation methods often lead to high holding costs and may have some practical drawbacks. We compare two customer differentiation policies: stock reservation and pipeline stock priority for high priority customers. We derive exact analytical expressions of the waiting time distribution of both types of customers for a stock reservation policy. We then provide accurate approximation methods for a pipeline stock priority policy. By comparison, we offer insights concerning which method should be used under different service level requirements.

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, this service level enables providers to engage customers in customer-focused performance metrics concerning costs associated with lost hours awaiting service parts (Cohen et al. 2006). Often, these costs are not linear in time but increase considerably after a certain time window (Fritzsche and Lasch 2012).

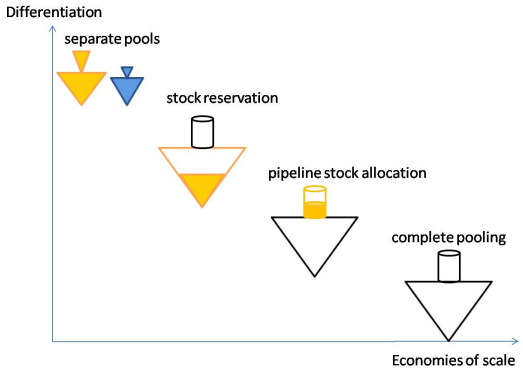
Different customers may have different service needs even when they own the same product. (Cohen et al. 2006). For example, when a mainframe computer in a stock exchange fails, the financial impact will be more severe than when a mainframe in a library goes down. In such situations, service providers 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 long response times and low fill rates correspond to the low priority customers.

The challenge to service providers is then 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, as shown in Figure 1. One way would be 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 propose to reserve a part of inventory for high priority customers and pool the rest (Deshpande et al. 2003). Despite the potential savings the stock reservation policies can offer, this method is not often observed in practice (Arslan et al. 2007). One of the reasons is that service engineers find it difficult to reserve a part for a future failure at a high priority customer while being asked to help a failure at a low priority customer (Alvarez et al. 2012).

As noted by many researchers (Dekker et al. 1998, Deshpande et al. 2003), a pipeline stock allocation policy may have significant impact on inventory system performance. Particularly, it affects the fill rate of high priority customers and the average length of stockouts of low priority customers (Dekker et al. 1998), thus influencing the total costs of the system (Deshpande et al. 2003).

Therefore, it is of great interest to investigate the customer-focused performance metrics under two different customer differentiation policies, namely, stock reservation for high priority customers

Figure 1 Customer Differentiation Policies



and pipeline stock priority for high priority customers. We contribute to the literature by providing exact analytical expressions of the customer-focused performance metrics for a stock reservation policy. We also contribute by providing accurate approximation methods for a pipeline stock priority policy.

The paper is organized as follows. In Section 2, we review the literature on customer differentiation policies. In Section 3, we present our model. In Section 4, we provide exact analytical expressions for the stock reservation policy. In Section 5, we give an approximation for the pipeline stock priority policy. In Section 6, we conduct extensive numerical experiments, to first validate our approximation method for the pipeline stock priority policy, then to compare it to the reservation policy. In the last section, we draw our conclusion and offer managerial insights.

2. Literature review

Veinott (1965) is one of the first to analyze an inventory model with several demand classes for a single product in a periodic review system. He shows that the base stock ordering policy is optimal if, in each period, on-hand inventory is used to satisfy high priority demand before any attempt to satisfy low priority demand. Furthermore, he proposes several useful variations of his model without carrying out the details of the analysis. One such variation is to use a critical level to ration the on-hand inventory among two demand classes. Topkis (1968) subsequently analyses the proposed critical level policy in a periodic review system, and proves that there exists an optimal, non-negative, critical level for each demand class.

Nahmias and Demmy (1981) analyze a critical level policy for a periodic review system and a continuous review (Q, R) inventory system with two demand classes. In the latter system, they assume no more than one outstanding order at any time, which allows them to approximate expected backorders and fill rates for both demand classes. They find that with the critical level policy, the overall fill rate for all demand classes is always lower than when no critical level is used. However, the fill rates for a high priority class are higher when the critical level policy is employed.

Ha (1997a) considers the inventory rationing problem in a make-to-stock production system with several demand classes and lost sales. He shows that the optimal policy is characterized by a critical stock level for each demand class. Ha (1997b) studies a similar problem but allows backordering. In his model, low priority backorders are cleared only when there is no backorder of the high priority class and a sufficient amount of stock has been built up in the system. For a Markovian model with exponentially distributed inter-arrival times and 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 low priority backorder), he shows that a base stock policy for the production decision and a dynamic critical level policy for rationing inventory are optimal.

Dekker et al. (1998) consider a model similar to the continuous review model discussed in Nahmias and Demmy (1981), but without the assumption on the number of outstanding orders. Their model also differs by considering an one-for-one inventory systems with deterministic lead time. They explore several ways of allocating the incoming pipeline stocks. The first allocation method gives priority first to backorders for critical demand then to backorders for non-critical demand and finally to stock reservation for critical demand. The second allocation method gives priority first to backorders for critical demand then to stock reservation for critical demand and finally to backorders for non-critical demand. For these methods, they provide approximations of fill rates for the high and low priority customers and show that the allocation methods have a significant impact on system performance.

Deshpande et al. (2003) consider the continuous review (Q, R) inventory system as in Nahmias and Demmy (1981), but without the assumption on the number of outstanding orders. They study a "threshold clearing" mechanism which allocates incoming pipeline stocks to all backorders, that were issued for all demand classes before the critical level hitting time, according to a first-come-first-served (FCFS) discipline. Only after clearing backorders, a"priority clearing" mechanism is applied to the remaining incoming pipeline stocks, which gives priority first to backorders for critical demand, then to stock reservation for critical demand and finally to backorders for non-critical demand. As a result, backorders originating from low priority customers can be cleared before backorders originating from high priority customers. This threshold clearing mechanism permits to derive expressions for the expected number of backorders for both classes, and to calculate the optimal ordering parameters and the optimal critical level. The parameters obtained from the "threshold clearing" scheme are then used in the "priority clearing" scheme. The authors show that this "hybrid" policy closely approximates the optimal "priority clearing" scheme.

Arslan et al. (2007) extend the work of Deshpande et al. (2003) to multiple demand classes by mapping this inventory system to an inventory system with multiple serial stages. Differing from previous work, for the allocation of incoming pipeline stocks between adjacent demand classes, they do not distinguish between backorders for higher priority demand, stock reservation for higher priority demand and backorders for lower priority demand, all of which are served according to a FCFS discipline. However, they establish the equivalence between the FCFS allocation rule and the threshold clearing process in Deshpande et al. (2003). They report that, though the FCFS allocation rule leads to lower fill rate than the priority allocation rule, the difference is quite small. They find that the cost in cases without the critical-level policy may be considerably higher than that in cases with critical-level policy.

In this paper, we compare three customer differentiation policies: separate pools of stocks for each customer class, stock reservation for high priority customers and pipeline stock priority for high priority customers. We first contribute to the literature on stock reservation policy, by providing exact analytical expressions of the customer-focused performance metrics, namely maximum response

time and level of guarantee. This customer-focused performance metrics distinguish themselves by their capability to best match customers' expectation. We then contribute to the literature on pipeline stock priority policy, by providing accurate approximation methods for the corresponding customer-focused performance metrics.

3. Model description

We consider a service parts inventory system with two demand classes, i.e., high priority or Gold customers and low priority or Silver customers. The two classes require different maximum response time guarantees. To provide differentiated services, we consider two differentiation policies: stock reservation for Gold customers and pipeline stock priority for Gold customers.

Assumptions and notations

Inventory control policy The inventory system is controlled by a one-for-one policy with base stock S under continuous review, i.e., every time a demand from either a Gold customer or a Silver customer occurs, a replenishment order is placed. The inventory system has no lost sales. All not satisfied demands are backlogged. This inventory control policy is commonly employed for service parts, which are characterized by high price and low demand rates (Sherbrooke 1968, Alfredsson and Verrijdt 1999).

Customer demand We assume that the demand processes corresponding to the two demand classes are independent Poisson processes. The demand rates of Gold and Silver customers will be denoted by λ_G and λ_Z respectively. Since the demand processes of Gold and Silver customers are independent Poisson processes, their superposition is also a Poisson process. As a result, the aggregated demand process is a Poisson process with rate $\lambda = \lambda_G + \lambda_Z$.

Lead times The lead times are assumed to be non-negative and deterministic, denoted by L.

Service performance matrices Customers can be differentiated by two service requirements: the response time and the service level within the response time i.e., the proportion of customers satisfied within the response time. Note that high priority customers require lower response time and in general the response time should be lower than the lead time.

We conclude this section by the list of parameters and Notations that will be used throughout the paper.

S = base stock

L = lead time

λ_G = arrival rate of Gold customers

λ_Z = arrival rate of Silver customers

λ = arrival rate of arbitrary customers

IL = inventory level

W_G^c = waiting time of a Gold customer in the stock reservation policy

W_Z^c = waiting time of a Silver customer in the stock reservation policy

W_G^p = waiting time of a Gold customer in the pipeline stock priority policy

W_Z^p = waiting time of a Silver customer in the pipeline stock priority policy

$\mathbf{Erl}_{n,\lambda}(\cdot)$ - the cumulative distribution function of an Erlang random variable with shape parameter n and rate λ

$\mathbf{Exp}_{\lambda}(\cdot)$ - the cumulative distribution function of an Exponential random variable with rate λ

$\mathbf{po}(\cdot;\beta)$ - 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

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

$X \sim F(\cdot)$ - the random variable X has the cumulative distribution $F(\cdot)$

$f * g$ - the convolution of the functions f and g.

4. Customer differentiation via stock reservation

In this section, we analyze the model with stock reservation policy, also known as critical level policy. We denote the critical level by K. In this inventory system, Silver customers are not served

any more once the inventory level reaches or falls below K. We consider the pipeline stock allocation method in Arslan et al. (2007), also known as first-come-first-served (FCFS) clearing mechanism, where shortfall for Gold customers and Silver customers' backorders are cleared according to a FCFS discipline. The shortfall for Gold customers is defined as the amount of inventory to clear all Gold backorders and to restore the on-hand inventory level to the critical inventory level.

For illustration purpose, consider a one-for-one inventory system with base stock $S = 3$ and critical level $K = 1$. We look at this system right after the on hand inventory has become 1, hence there are 2 items in pipeline and no customer is waiting. Since $K = 1$, the item on stock can only be given to a Gold customer. Suppose that three customers, with Gold, Silver and Gold priorities, arrive before the first pipeline item joins the stock. In this case, the first Gold customer will get the item on stock, thus causing a shortfall. Since the shortfall was registered before the arrival of the Silver customer, the first pipeline item will be used to restore the shortfall, and raise the on hand inventory level to 1. Since $K = 1$, the item on stock is reserved for Gold customers and cannot be given to the waiting Silver customer. The Silver customer is backordered and the pipeline item will be given to the second Gold customer, and a new shortfall will be registered. Since the Silver customer arrived before the second shortfall has been registered, he will get the second pipeline item. Now suppose that the three customers had Silver, Gold and Gold priorities. Since the item on stock is reserved for Gold customers, it will be given to the first Gold customer and a shortfall will be registered. However, since the Silver customer arrived before the shortfall was registered, the first item in pipeline will be given to the Silver customer.

The analysis of this model relies on the equivalence between a demand-class inventory system with a stock reservation policy (DCS) and the following serial stage inventory system (SSS) . For a proof of this equivalence, we refer to Arslan et al. (2007) . We present this equivalence for 2 demand classes. The (SSS) inventory system divides the on hand inventory into 2 stockpiles (stages), one for each demand class. The first stockpile operates with a continuous review base stock policy with base-stock level K and is replenished by the second stockpile with zero replenishment time. The second stockpile uses a continuous review base stock policy with base stock level $S - K$. The second

stockpile uses an outside supplier with leadtime L. The demand at the first stockpile follows a Poisson process with rate λ_G. Demand at second stage is divided into internal demand with rate λ_G, originating from the first class (Gold customers), and external demand, corresponding to Silver customers, with rate λ_Z. Internal demand corresponds to recovering of the reservation stock for Gold customers. At each stage, all internal and external demands that cannot be fulfilled from on hand inventory is backordered. Backorders are served in a FCFS manner.

Recall that in a classical continuous review, base stock model, with base stock level S and one class of customers, a pipeline item is given to the $S-$th customer arriving after this item was ordered. Consider now our model with two demand classes and stock reservation policy and the equivalent (SSS) model. Tag a pipeline item right after the time $\tilde{t}$ when it is ordered. We will refer to customers/demand requests after time $\tilde{t}$ as future customers/demand requests. The tagged pipeline item will satisfy the $S-K$-th (internal and external) future demand request at the second stockpile, say at $\tilde{t}+Y_{S-K,\lambda}$. Specifically, suppose the demand is external, thus corresponds to a demand from a Silver customer. In this case, the tagged item is given to the $S-K$th customer arriving after $\tilde{t}$. If the demand is internal, thus caused by a shortfall, the tagged item will be used to restore the reservation stock for Gold customers, i.e., raise the inventory level of the first stockpile to K. In this case, the tagged item will be given to the $K-$th Gold customer arriving after $\tilde{t}+Y_{S-K,\lambda}$.

In Proposition 1 we derive the distribution of the waiting time for a Silver customer, while in Proposition 2 we derive the distribution of a Gold customer. The proofs can be found in Appendix A.

PROPOSITION 1. *In a single echelon system with critical level policy, where pipeline stocks are allocated according to a FCFS clearing mechanism, the waiting time distribution for a Silver customer is given by*

$$P(W_Z^c \leq t) = 1 - \boldsymbol{Erl}_{S-K,\lambda}(L-t) = \boldsymbol{Po}(S-K-1;\lambda(L-t)). \tag{1}$$

PROPOSITION 2. *In a single echelon system with critical level policy, where pipeline stocks are allocated according to a FCFS clearing mechanism, the waiting time distribution for a Gold customer is given by*

$$P(W_G^c \leq t) = 1 - \boldsymbol{Erl}_{S-K,\lambda} * \boldsymbol{Erl}_{K,\lambda_G}(L-t). \tag{2}$$

Note that for $K = 0$, the model with reservation reduces to a model without reservation where all customers are served according to the FCFS rule. In this case, based on the fact that $\mathbf{Erl}_{0,\lambda}(\cdot) = 1$, $1 - \mathbf{Erl}_{S,\lambda}(L) = \mathbf{Po}(S-1, \lambda L)$ the results in Proposition 1 and Proposition 2 , coincide with the result for a regular FCFS inventory system (see Proposition 1 in Yang et al. (2012)).

5. Customer differentiation via pipeline stock priority

Next we analyze the model where Gold customers have priority over Silver customers for pipeline stock when there is no stock on hand, and served in order of arrival otherwise. We assume no stock reservation for Gold customers. For illustration purpose, consider a one-for-one inventory system with base stock $S = 3$ just before the arrival of a pipeline item in stock. Suppose that at this moment, there is zero on-hand inventory, one Gold and one Silver backorder. Since Gold customers have priority over Silver customers with respect to pipeline items, the next arriving pipeline stock will be allocated to clear the Gold backorder, even if the Gold customer has arrived after the Silver one.

For this model, we propose simple and accurate approximations for the service levels of the two demand classes. The performance of these heuristics is studied in Section 6. We use superscript p to distinguish the performance measures for this differentiation method.

5.1. Approximate waiting time distribution of a Gold customer

The approximation relies on the well know equivalence between an one-for-one inventory system with Poisson demands and an $M/G/\infty$ queue. Next we are breifly reviewing this equivalence and a few classical results that rely on it. The arrivals in the $M/G/\infty$ queue are the orders placed at the arrival of a customer in the one-for-one inventory system with Poisson demands. The service time

is equivalent to the lead time. Note that the priorities do not affect the arrival of ordered items via pipeline; they only affect the order in which waiting customers are served after the items arrive in stock. By Palm's theorem, the probability that a customer sees at arrival n items in pipeline is $\mathbf{po}(n, \lambda L)$.

Brumelle (1978) shows that if an arrival at an $M/G/\infty$ queue finds n busy servers, the remaining service times are distributed as n independent random variables with distribution $H^*(x) = \frac{\int_0^x P(S>t)dt}{E(S)}$, where S is the service time. In the case of constant service time L, $H^*(x) = \frac{x}{L}$ for $x \in [0, L]$. Assume that the items in pipeline are numbered in the order they arrive into stock. The time till the arrival of the $k+1$-th pipeline stock is thus given by $U_{(k+1)}$, the $k+1$-th order statistics of n uniformly distributed random variables on $[0, L]$ and thus follows a $\mathbf{Beta}(\frac{\cdot}{L}; k+1, n-k)$ distribution.

Tag a Gold customer at his arrival. With probability $\mathbf{po}(n, \lambda L)$, he sees n items in pipeline at arrival.

If the tagged customer sees stock on hand, i.e., if he sees $n \leq S-1$ items in the pipeline, he is immediately served. In this case,

$$P(W_G^p = 0) = \mathbf{Po}(S-1; \lambda L). \tag{3}$$

If $n \geq S$, the tagged customer sees at his arrival $n-S$ customers waiting. Due to the priority rule, the Gold customer only has to wait till all the other waiting Gold customers are served. Our approximation relies on the following assumption:

Assumption 1 A waiting customer is a Gold customer with probability q_G, independent of other customers.

Assume for the moment that the value of q_G is known. Then, with probability $\mathbf{bin}(k; n-S, q_G) = \binom{n-S}{k} q_G^k (1-q_G)^{n-S-k}$, k out of the $n-S$ waiting customers are Gold customers. In this case, the Gold customer with get $k+1$-th pipeline item and his waiting time is equal to the residual lead time of this item. Based on Brumelle (1978), the residual lead time of $k+1$-th item follows a $\mathbf{Beta}(\frac{\cdot}{L}; k+1, n-k)$ distribution.

Combining the above ideas, we obtain the following approximation for the distribution of the waiting time of a Gold customer.

$$P(0 < W_G^p \leq t) \simeq \sum_{n=S}^{\infty} \mathbf{po}(n; \lambda L) \sum_{k=0}^{n-S} \mathbf{bin}\,(k; n-S, q_G)\, \mathbf{Beta}(\frac{t}{L}; k+1, n-k) \tag{4}$$

Note that for a system with one demand class, where $q_G = 1$, the approximation is exact, and coincides with the results in Proposition 1 in Yang et al. (2012).

Approximating q_G Since the two types of customers arrive according to independent Poisson processes, the probability that a customer is a Gold customer is $\frac{\lambda_G}{\lambda}$. However, since Gold customers have priority, q_G must be smaller than $\frac{\lambda_G}{\lambda}$. We approximate the probability q_G by $\left(\frac{\lambda_G}{\lambda}\right)^2$. This choice can be justified as follows. Suppose that each pipeline item is given to the customer who issued the order L time periods before. Then, a pipeline item would be given to a Gold and Silver customer with probability $\frac{\lambda_G}{\lambda}$ and $\frac{\lambda_Z}{\lambda}$, respectively. We assume next that when a Gold customer gets priority, will actually receive an item dedicated to a Silver customer. The probability that a customer is a Gold customer and gets an item dedicated to a Silver customer is thus equal to $\frac{\lambda_G}{\lambda}\frac{\lambda_Z}{\lambda}$. Finally, we approximate the probability that a waiting customer is a Gold customer by the probability that he is a Gold customer, but will not make use of his priority, which means that he will get an item dedicated to a Gold customer. This happens with probability $\frac{\lambda_G}{\lambda} - \frac{\lambda_G \lambda_Z}{\lambda^2} = \left(\frac{\lambda_G}{\lambda}\right)^2$.

In Section 6 the accuracy of this approximation is tested for different parameter settings.

5.2. Approximate waiting time distribution of a Silver customer

Tag a Silver customer at his arrival, say time t. We take the arrival of the Silver customer as time reference. Assume that items in pipeline are numbered in increasing order of the residual lead time. As for the Gold customers, we can use Palm's theorem to argue that

$$P(W_Z^p = 0) = \mathbf{Po}(S-1, \lambda L).$$

Suppose the tagged customer sees $n \geq S$ items in pipeline, or, $n - S$ customers waiting in front of him. In a system without priorities, the Silver customer would get $n - S + 1$-th item in pipeline. However, in a system with priorities, the Silver customer gets item $n - S + 1 + k$, if k Gold customers

arrive between t and the arrival of item $n-S+1$ in stock. A complicating factor in calculating the waiting time distribution of Silver customers is that the residual lead times of the items in pipeline and the number of Gold customers that arrive between t and the service of the tagged Silver customer are dependent variables. In order to simplify the calculations we make the following assumption.

Assumption 2 A Silver customer who upon arrival sees n items in pipeline, $n \geq S$, will get item $n-S+1+k$ with probability $p_{n,k}$, independent of the residual lead time of the items in pipeline.

Assume further that the tagged Silver customer gets item $n-S+1+k$ and that the values of $p_{n,k}$ are known.

Suppose that $k \leq S-1$. In this case, the Silver customer gets an item that arrives into stock before item $n+1$, hence an item he seen in pipeline at arrival. Therefore, his waiting time will be smaller than L with probability 1. As argued in Section 5.1, the residual lead time of l-th item in pipeline follows a $\mathbf{Beta}(\frac{\cdot}{L}, l, n-l+1)$ distribution.

If $k=S$, the tagged customer gets item $n+1$, hence the item ordered at his arrival. In this case, the waiting time is equal to L.

Assume now that $k \geq S+1$. In this case, the tagged customer will get the pipeline item with index $n-S+k+1 > n+1$ and thus his waiting time will be larger than L. Note that since items $n+1, ..., n-S+1+k$ were ordered after the arrival of the Silver customer, the times between their arrival in stock are exponentially distributed with rate λ. Hence the waiting time of the Silver customer is equal to $L+A$, where $A = \mathbf{Erl}_{k-S,\lambda}(\cdot)$.

Based on the ideas above we propose the following approximation for the distribution of W_Z:

$$P(W_Z^p \leq t) = P(W_Z = 0) + \sum_{n=S}^{\infty} \mathbf{po}(n, \lambda L) \sum_{k=0}^{\infty} p_{n,k} P(C_{n,k} \leq t), \tag{5}$$

where

$$P(C_{n,k} \leq t) = \begin{cases} \mathbf{Beta}(\frac{t}{L}; n-S+k+1, S-k) \text{ for } k \leq S-1, t < L, \\ 0 \text{ for } k = S, t < L, \\ 1 \text{ for } k \leq S, t \geq L, \\ \mathbf{Erl}_{k-S,\lambda}(t-L) \text{ for } k \geq S+1. \end{cases}$$

Approximating $p_{n,k}$ In the next two lemma's we propose approximations for the probabilities $p_{n,k}$. Recall that the residual lead times of the pipeline items follow a Beta distribution. In order to simplify the calculations, the lemma's make use of the following assumption.

Assumption 3 If a customer sees n items in pipeline at his arrival, the time intervals between the arrival of $k-$th and $k+1$-th item in stock, $1 \leq k \leq n$, are independent and exponentially distributed with mean $\frac{L}{n+1}$.

Under this assumption, the residual lead times become Erlang distributed, which leads to closed formulas for the probabilities $p_{n,k}$. The proofs of the lemma's, together with some preliminary results, can be found in Appendix B.

LEMMA 1. *Under Assumption 3, the probability* $p_{n,k}$ *that a Silver customer who sees* $n \geq S$ *items in pipeline gets item* $n-S+1+k$*, for* $1 \leq k \leq S$*, is given by*

$$p_{n,0} = \left(\frac{\beta}{\lambda_G+\beta}\right)^{n-S+1},$$

for $k=0$ *and by*

$$p_{n,k} = \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{k}\left(\frac{\beta}{\lambda_G+\beta}\right)^{n-S+k+1}\sum_{a=1}^{k}\frac{a}{k}\binom{n-S+a}{a}\binom{2k-a-1}{k-a},$$

where $\beta = \frac{n+1}{L}$.

Next lemma gives an approximation of $p_{n,k}$ for $k \geq S+1$.

LEMMA 2. *Under Assumption 3, the probability* $p_{n,k}$ *that a Silver customer who sees* $n \geq S$ *items in pipeline gets item* $n-S+1+k$*, with* $k \geq S+1$*, is given by*

$$p_{n,k} = \left(\frac{\lambda}{\lambda_G+\lambda}\right)^{k-S}\left(\frac{\beta}{\lambda_G+\beta}\right)^{n+1}\sum_{a=1}^{k}\binom{n-S+a}{a}\sum_{B=\max\{S+1-a,0\}}^{k-a}\left(\frac{\lambda_G}{\lambda_G+\lambda}\right)^{k-B}\left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{B}q_k(a,B),$$

with

$$q_k(a,B) = \begin{cases} \frac{a+B-S}{k-S}\binom{2k-a-B-S-1}{k-S-1}\left(\binom{B+S-1}{B}-\binom{B+S-1}{a+B}\right) & \text{for } a < S \\ \frac{a+B-S}{k-S}\binom{2k-a-B-S-1}{k-S-1}\binom{B+S-1}{B}, & \text{for } a \geq S. \end{cases}$$

The accuracy of the proposed approximation and a comparison between stock reservation and pipeline stock priority are discussed in Section 6 .

6. Numerical experiments

In this section, we first validate the approximations proposed in Section 5. Then, via numerical experiments we compare the two methods of differentiation, namely stock reservation and pipeline stock priority. The following notations will be used throughout this section.

τ - response time for Gold customer

ξ - response time for Silver customers

SL_G^c- service level within response time τ for Gold customers under the stock reservation policy

SL_G^p- service level within response time τ for Gold customers under the pipeline stock priority policy

To calculate the service levels $SL_Z^c = P(W_Z^c \leq \xi)$ and $SL_G^c = P(W_G^c \leq \tau)$ for the stock reservation policy, we use (1) and (2), respectively. The service levels $SL_G^p = P(W_G^p \leq \tau)$ and $SL_Z^p = P(W_Z^p \leq \tau)$ under the pipeline priority policy will be approximated by using (4) and (5).

The test bed 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. We vary the base stock S between 7 and 14 and the proportion of Gold and Silver customers, $\frac{\lambda_G}{\lambda_Z} \in \{0.5, 1, 2\}$.

To study the quality of the approximation proposed in Section 5 we evaluate the service level of Gold customers, $SL_G^p = P(W_G^p \leq \tau)$ for $\tau \in \{0.25, 0.4, 0.5, 0.6, 0.75\}$ with $q_G = \left(\frac{\lambda_G}{\lambda}\right)^2$. For each value of τ, the approximated service levels and the difference between them and the service levels obtained by simulation are reported in columns 5 to 16 in Table 1. As the results in Table 1 show, in the cases we studied, the approximation is very accurate, the average absolute error being 0.38%. The maximum absolute error, of 3.49% is obtained for case 1, which is characterized by a low S and a service level below 50%, which is rarely desirable in practice for the highest priority class. Moreover, the quality of the approximation improves with an increase in S. For lower values of S, i.e. $7 \leq S \leq 10$, the approximation is most accurate for the cases where the proportion of Silver customers is low, i.e. $\lambda_G = 1.5, \lambda_Z = 0.75$ and least accurate for high proportion of Silver customers, i.e., $\lambda_G = 0.75, \lambda_Z = 1.5$. However, when S increases, i.e. $12 \leq S \leq 14$, the proportion of Gold and

Silver customers seems to affect less the quality of the approximation, as the cases corresponding to $\lambda_G = 1.5, \lambda_Z = 0.75$ and $\lambda_G = 0.75, \lambda_Z = 1.5$ indicate. In these cases, the error seems to be most influenced by the total load, as the higher errors for the cases with $\lambda_G = \lambda_Z = 1.5$ suggest. Over the tested values of τ, the highest average absolute error (equal to 0.48%) is obtained for $\tau = 0.25$. Note that for this value of τ, the approximation understimates the real service level for the cases with higher proportion of Gold customers (i.e., $\lambda_G = 1.5, \lambda_Z = 0.75$ and $\lambda_G = 1.5, \lambda_Z = 1.5$), and slightly overestimates when the proportion of Gold customers is low (i.e. $\lambda_G = 0.75, \lambda_Z = 1.5$). In general, the approximation seems to underestimate the real service level for low values of τ and low values of S, while for the other cases, it slightly overestimates it.

Table 1 Performance of the approximation method for the service level of Gold customers in the pipeline stock priority policy

	Inputs			$\tau = 0.25$		$\tau = 0.4$		$\tau = 0.5$		$\tau = 0.6$		$\tau = 0.75$		$\tau = 1$	
Cases	S	λ_G	λ_Z	SL_G^p	$Error$[a]	SL_G^p	$Error$[a]	SL_G^p	$Error$[a]	SL_G^p	$Error$[a]	SL_G^p	$Error$[a]	SL_G^p	$Error$[a]
1	7	1.5	1.5	48.15%	-3.49%	61.78%	-2.36%	69.46%	-1.42%	76.02%	-0.69%	83.86%	0.34%	92.43%	1.38%
2	7	1.5	0.75	64.47%	-1.53%	72.53%	-0.80%	77.24%	-0.64%	81.42%	-0.17%	86.68%	0.37%	92.99%	0.82%
3	7	0.75	1.5	72.31%	0.76%	81.48%	1.21%	86.04%	1.56%	89.62%	1.69%	93.51%	1.35%	97.25%	0.94%
4	8	1.5	1.5	59.08%	-2.04%	71.25%	-0.74%	77.77%	-0.12%	83.13%	0.43%	89.23%	0.86%	95.42%	1.12%
5	8	1.5	0.75	76.69%	-0.76%	82.88%	-0.27%	86.31%	0.01%	89.23%	0.31%	92.72%	0.43%	96.55%	0.53%
6	8	0.75	1.5	81.75%	0.64%	88.37%	0.84%	91.53%	1.01%	93.91%	0.99%	96.40%	0.95%	98.63%	0.51%
7	9	1.5	1.5	69.60%	-0.54%	79.69%	0.10%	84.83%	0.46%	88.89%	0.84%	93.30%	0.94%	97.42%	0.86%
8	9	1.5	0.75	85.92%	-0.23%	90.18%	-0.04%	92.44%	0.07%	94.27%	0.29%	96.35%	0.35%	98.45%	0.31%
9	9	0.75	1.5	88.92%	0.27%	93.29%	0.52%	95.28%	0.50%	96.73%	0.63%	98.17%	0.46%	99.37%	0.24%
10	10	1.5	1.5	78.80%	-0.17%	86.55%	0.35%	90.31%	0.64%	93.16%	0.75%	96.10%	0.76%	98.65%	0.45%
11	10	1.5	0.75	92.14%	-0.08%	94.81%	0.21%	96.15%	0.14%	97.19%	0.17%	98.32%	0.21%	99.36%	0.16%
12	10	0.75	1.5	93.79%	0.32%	96.43%	0.32%	97.58%	0.27%	98.38%	0.31%	99.15%	0.24%	99.74%	0.09%
13	11	1.5	1.5	86.12%	0.10%	91.65%	0.46%	94.20%	0.53%	96.06%	0.67%	97.88%	0.46%	99.34%	0.30%
14	11	1.5	0.75	95.93%	-0.11%	97.46%	0.03%	98.18%	0.08%	98.73%	0.11%	99.28%	0.11%	99.76%	0.06%
15	11	0.75	1.5	96.78%	0.12%	98.24%	0.21%	98.85%	0.15%	99.26%	0.13%	99.63%	0.10%	99.90%	0.04%
16	12	1.5	1.5	91.46%	0.15%	95.13%	0.35%	96.74%	0.36%	97.87%	0.38%	98.92%	0.32%	99.70%	0.15%
17	12	1.5	0.75	98.04%	-0.02%	98.84%	0.08%	99.20%	0.05%	99.46%	0.05%	99.72%	0.03%	99.91%	0.02%
18	12	0.75	1.5	98.44%	0.09%	99.19%	0.07%	99.49%	0.08%	99.69%	0.09%	99.85%	0.07%	99.96%	0.01%
19	13	1.5	1.5	95.06%	0.15%	97.33%	0.27%	98.28%	0.21%	98.92%	0.20%	99.49%	0.15%	99.87%	0.07%
20	13	1.5	0.75	99.12%	0.01%	99.51%	0.02%	99.67%	0.03%	99.79%	0.03%	99.90%	0.02%	99.97%	0.01%
21	13	0.75	1.5	99.30%	0.02%	99.66%	0.01%	99.79%	0.02%	99.88%	0.03%	99.95%	0.02%	99.99%	0.00%
22	14	1.5	1.5	97.31%	0.13%	98.62%	0.13%	99.15%	0.08%	99.48%	0.10%	99.77%	0.07%	99.95%	0.02%
23	14	1.5	0.75	99.63%	0.01%	99.80%	0.01%	99.87%	0.00%	99.92%	0.01%	99.96%	0.01%	99.99%	0.00%
24	14	0.75	1.5	99.70%	0.00%	99.86%	0.01%	99.92%	0.02%	99.95%	0.01%	99.98%	0.00%	100.00%	0.00%

[a] $Error = SL_G^p -$ service level of Gold customers obtained by simulation.

Table 2 contains the comparison between the approximation of the service level for Silver customers proposed in Section 5 and simulations. Columns 5 -16 report the approximated service levels $SL_Z^p = P(W_Z^p \leq \xi)$ for $\xi \in \{0.25, 0.4, 0.5, 0.6, 0.75, 1\}$ and the difference between them and the service levels obtained by simulation.

The average absolute error of the approximation is less than 1.9% with a maximum absolute error of 6.68% in Case 3, when $S = 7$ and $\xi = 0.5$. The approximation overestimated the real service level for Silver customers in all tested cases. For each proportion of Gold and Silver customers,

Table 2 Performance of the approximation method for the service level of Silver customers in the pipeline stock priority policy.

Inputs				$\zeta = 0.25$		$\zeta = 0.4$		$\zeta = 0.5$		$\zeta = 0.6$		$\zeta = 0.75$		$\zeta = 1$	
Cases	S	λ_G	λ_Z	SL_Z^p	$Error^a$	SL_Z^p	$Error^a$	SL_Z^p	$Error^a$	SL_Z^p	$Error^a$	SL_Z^p	$Error^a$	SL_Z^p	$Error^a$
1	7	1.5	1.5	29.29%	0.53%	38.38%	5.78%	41.26%	5.98%	43.83%	6.36%	47.78%	6.26%	53.24%	5.30%
2	7	1.5	0.75	58.01%	0.39%	65.65%	4.32%	68.24%	4.54%	70.45%	4.61%	73.40%	4.47%	77.21%	3.80%
3	7	0.75	1.5	60.02%	1.65%	65.07%	2.40%	72.15%	6.68%	74.81%	6.48%	78.22%	6.14%	83.17%	4.70%
4	8	1.5	1.5	42.75%	0.53%	52.29%	6.09%	55.31%	6.00%	57.96%	5.86%	61.86%	5.74%	67.08%	4.24%
5	8	1.5	0.75	72.00%	0.49%	78.13%	3.17%	80.20%	3.28%	81.93%	3.20%	84.16%	3.01%	86.92%	2.05%
6	8	0.75	1.5	73.63%	1.49%	80.70%	4.84%	83.10%	4.69%	85.11%	4.42%	87.74%	4.25%	90.98%	3.00%
7	9	1.5	1.5	56.53%	0.82%	65.43%	5.16%	68.24%	5.37%	70.88%	5.29%	74.07%	4.64%	78.44%	3.10%
8	9	1.5	0.75	82.83%	0.29%	87.19%	2.10%	88.65%	2.23%	89.89%	2.22%	91.34%	1.93%	93.09%	1.01%
9	9	0.75	1.5	84.00%	1.01%	88.92%	3.04%	90.57%	2.92%	91.91%	2.77%	93.60%	2.42%	95.56%	1.58%
10	10	1.5	1.5	69.10%	0.82%	76.57%	4.40%	78.92%	3.89%	81.05%	3.99%	83.56%	3.55%	86.84%	2.04%
11	10	1.5	0.75	90.29%	0.13%	93.07%	1.26%	94.00%	1.31%	94.76%	1.16%	95.63%	0.86%	96.63%	0.42%
12	10	0.75	1.5	91.04%	0.63%	94.12%	1.93%	95.14%	1.74%	95.98%	1.73%	96.92%	1.22%	97.98%	0.67%
13	11	1.5	1.5	81.54%	2.90%	85.14%	3.14%	86.99%	3.02%	88.47%	2.74%	90.27%	2.09%	92.49%	1.05%
14	11	1.5	0.75	95.40%	0.68%	96.53%	0.76%	97.07%	0.74%	97.49%	0.56%	97.95%	0.34%	98.47%	0.21%
15	11	0.75	1.5	95.35%	0.29%	97.11%	1.04%	97.68%	0.92%	98.13%	0.78%	98.62%	0.50%	99.15%	0.27%
16	12	1.5	1.5	88.59%	2.06%	91.16%	2.01%	92.44%	1.88%	93.43%	1.78%	94.60%	1.30%	96.73%	1.28%
17	12	1.5	0.75	97.77%	0.27%	98.38%	0.33%	98.67%	0.37%	98.88%	0.23%	99.11%	0.20%	99.45%	0.17%
18	12	0.75	1.5	97.76%	0.26%	98.68%	0.50%	98.97%	0.43%	99.19%	0.31%	99.43%	0.26%	99.66%	0.08%
19	13	1.5	1.5	93.37%	1.09%	95.06%	1.33%	95.87%	1.26%	96.48%	0.99%	97.18%	0.70%	98.36%	0.51%
20	13	1.5	0.75	98.99%	0.13%	99.30%	0.15%	99.43%	0.12%	99.53%	0.12%	99.63%	0.07%	99.78%	0.03%
21	13	0.75	1.5	99.11%	0.20%	99.43%	0.21%	99.57%	0.18%	99.67%	0.17%	99.78%	0.10%	99.87%	0.03%
22	14	1.5	1.5	96.37%	0.67%	97.41%	0.71%	97.87%	0.56%	98.23%	0.55%	98.61%	0.33%	99.21%	0.25%
23	14	1.5	0.75	99.57%	0.09%	99.71%	0.06%	99.77%	0.07%	99.82%	0.03%	99.86%	0.02%	99.91%	0.00%
24	14	0.75	1.5	99.63%	0.11%	99.77%	0.10%	99.83%	0.08%	99.88%	0.05%	99.92%	0.04%	99.95%	0.01%

[a] $Error = SL_Z^p -$ service level for Silver customers obtained by simulation.

the average absolute errors (over the different values of ξ) decrease as the value of S increases. As for Gold customers, for each value of S, the lowest errors are obtained for the cases with a low proportion of Silver customers, i.e., $\lambda_G = 1.5, \lambda_Z = 0.75$. As the value of S increases, the proportion of Gold and Silver customers does not seem to play an important role. For all values of S, the highest errors are obtained for the cases with high load, namely $\lambda_G = \lambda_Z = 1.5$. Note that for most of these cases, the service level for Silver customers is below 70%. The approximation seems to work well for low and large response times ($\xi = 0.25$ and $\xi = 1$), with an average absolute error (over 24 cases) below 1.5%. The largest average absolute error, of 2.43% is registered for $\xi = 0.5$.

Tables 3 - 6 compare the performance of the stock reservation policy with various critical levels and the pipeline stock priority policy with different response times for Gold and Silver customers, given the same investment in base stocks. Each table corresponds to different response times τ and ξ. In all tables we vary S between 11 and 14, since for these cases the error of our approximations is below 3.5% and does not affect the comparison. Columns 5 and 7 in all tables show the service levels for Gold and Silver customers using stock reservation with $K = 2$, columns 9 and 11 report the service levels with $K = 4$, and columns 13 and 15 report the service levels with $K = 6$. As expected, the numerical results indicate that a stock reservation policy with higher critical level leads to higher service level for Gold but lower service level for Silver customers. For example, the

Table 3 Comparison between stock reservation policy and pipeline stock priority policy, $L = 3$**,** $\tau = 0.25$**,** $\zeta = 0.5$**.**

Inputs				$K=2$				$K=4$				$K=6$			
Cases	S	λ_G	λ_Z	SL_G^c	Δ^a	SL_Z^c	Δ^b	SL_G^c	Δ^a	SL_Z^c	Δ^b	SL_G^c	Δ^a	SL_Z^c	Δ^b
1	11	1.5	1.5	89.28%	3.17%	66.20%	-20.80%	94.72%	8.61%	37.82%	-49.18%	97.48%	11.36%	13.21%	-73.79%
2	11	1.5	0.75	96.81%	0.87%	88.34%	-8.73%	98.01%	2.08%	66.63%	-30.44%	98.77%	2.84%	33.84%	-63.23%
3	11	0.75	1.5	98.81%	2.04%	88.34%	-9.34%	99.73%	2.96%	66.63%	-31.05%	99.94%	3.17%	33.84%	-63.84%
4	12	1.5	1.5	93.64%	2.17%	77.64%	-14.80%	97.00%	5.53%	52.46%	-39.97%	98.62%	7.16%	24.14%	-68.29%
5	12	1.5	0.75	98.49%	0.45%	93.95%	-4.72%	99.08%	1.04%	79.38%	-19.29%	99.44%	1.40%	50.76%	-47.90%
6	12	0.75	1.5	99.46%	1.01%	93.95%	-5.03%	99.88%	1.44%	79.38%	-19.60%	99.98%	1.53%	50.76%	-48.21%
7	13	1.5	1.5	96.43%	1.37%	86.22%	-9.65%	98.38%	3.32%	66.20%	-29.67%	99.28%	4.22%	37.82%	-58.05%
8	13	1.5	0.75	99.33%	0.21%	97.10%	-2.33%	99.60%	0.48%	88.34%	-11.09%	99.76%	0.64%	66.63%	-32.80%
9	13	0.75	1.5	99.77%	0.47%	97.10%	-2.47%	99.95%	0.65%	88.34%	-11.23%	99.99%	0.69%	66.63%	-32.95%
10	14	1.5	1.5	98.11%	0.80%	92.08%	-5.80%	99.17%	1.86%	77.64%	-20.23%	99.64%	2.34%	52.46%	-45.41%
11	14	1.5	0.75	99.72%	0.09%	98.71%	-1.06%	99.84%	0.21%	93.95%	-5.83%	99.91%	0.28%	79.38%	-20.39%
12	14	0.75	1.5	99.91%	0.20%	98.71%	-1.12%	99.98%	0.28%	93.95%	-5.89%	100.00%	0.29%	79.38%	-20.45%

[a] $\Delta = SL_G^c - SL_G^p$ [b] $\Delta = SL_Z^c - SL_Z^p$

service level for Gold customers increases in Case 1 in Table 3 from 89.28% ($K = 2$) to 97.48% ($K = 6$) while the service level for Silver customers decreases from 66.20% ($K = 2$) to 13.21% ($K = 6$). On the other hand, the pipeline stock priority policy can better balance the services to Gold and Silver customers after differentiation by response times. For example, in Case 1, $K = 2$, in Table 3, the service level for Gold customers is 86.12% while the service level for Silver customers is 87.16% when using the pipeline stock priority policy. A similar behaviour can be noticed by analysing the results in Tables 4-6.

Columns 6, 10, 14 in Tables 3 - 6 show the difference in service level for Gold customers between the stock reservation policy with various critical levels and the pipeline stock priority policy. The experiments suggest that, in terms of service level for Gold customers, the stock reservation policy has an advantage of up to 11.36% in Case 1, $K = 6$ in Table 3. The reason is that the stock reservation policy gives Gold customers access to more resources all the time, whereas the pipeline stock priority policy only does that occasionally. As we can see in Cases 1, 4, 7, and 10 in all the tables, the advantage is highest when the arrival rate of Silver customers is high. This effect decreases as the base stock S increases.

However, as shown in Columns 8, 12, 16, in all tables, the stock reservation policy offers much lower service levels to Silver customers comparing to the pipeline stock priority policy. The difference can raise up to to 73.95% in Case 1, $K = 6$, in Table 3. In all tables, the differences between the service levels offered by the two policies are higher in the cases with higher arrival rates and lower S values.

Table 4 Comparison between stock reservation policy and pipeline stock priority policy, $L = 3$, $\tau = 0.25$, $\zeta = 0.75$.

Inputs				$K = 2$				$K = 4$				$K = 6$			
Cases	S	λ_G	λ_Z	SL_G^c	Δ^a	SL_Z^c	Δ^b	SL_G^c	Δ^a	SL_Z^c	Δ^b	SL_G^c	Δ^a	SL_Z^c	Δ^b
1	11	1.5	1.5	89.28%	3.17%	76.11%	-14.16%	94.72%	8.61%	48.76%	-41.51%	97.48%	11.36%	19.70%	-70.56%
2	11	1.5	0.75	96.81%	0.87%	92.77%	-5.18%	98.01%	2.08%	75.30%	-22.65%	98.77%	2.84%	42.96%	-54.99%
3	11	0.75	1.5	98.81%	2.04%	92.77%	-5.85%	99.73%	2.96%	75.30%	-23.32%	99.94%	3.17%	42.96%	-55.66%
4	12	1.5	1.5	93.64%	2.17%	85.49%	-9.11%	97.00%	5.53%	63.59%	-31.01%	98.62%	7.16%	33.38%	-61.22%
5	12	1.5	0.75	98.49%	0.45%	96.58%	-2.52%	99.08%	1.04%	86.00%	-13.10%	99.44%	1.40%	60.50%	-38.61%
6	12	0.75	1.5	99.46%	1.01%	96.58%	-2.84%	99.88%	1.44%	86.00%	-13.43%	99.98%	1.53%	60.50%	-38.93%
7	13	1.5	1.5	96.43%	1.37%	91.83%	-5.35%	98.38%	3.32%	76.11%	-21.08%	99.28%	4.22%	48.76%	-48.42%
8	13	1.5	0.75	99.33%	0.21%	98.51%	-1.12%	99.60%	0.48%	92.77%	-6.86%	99.76%	0.64%	75.30%	-24.34%
9	13	0.75	1.5	99.77%	0.47%	98.51%	-1.26%	99.95%	0.65%	92.77%	-7.00%	99.99%	0.69%	75.30%	-24.48%
10	14	1.5	1.5	98.11%	0.80%	95.71%	-2.90%	99.17%	1.86%	85.49%	-13.12%	99.64%	2.34%	63.59%	-35.02%
11	14	1.5	0.75	99.72%	0.09%	99.40%	-0.46%	99.84%	0.21%	96.58%	-3.27%	99.91%	0.28%	86.00%	-13.86%
12	14	0.75	1.5	99.91%	0.20%	99.40%	-0.52%	99.98%	0.28%	96.58%	-3.33%	100.00%	0.29%	86.00%	-13.92%

[a] $\Delta = SL_G^c - SL_G^p$ [b] $\Delta = SL_Z^c - SL_Z^p$

Table 5 Comparison between stock reservation policy and pipeline stock priority policy, $L = 3$, $\tau = 0.5$, $\zeta = 0.75$.

Inputs				$K = 2$				$K = 4$				$K = 6$			
Cases	S	λ_G	λ_Z	SL_G^c	Δ^a	SL_Z^c	Δ^b	SL_G^c	Δ^a	SL_Z^c	Δ^b	SL_G^c	Δ^a	SL_Z^c	Δ^b
1	11	1.5	1.5	93.40%	-0.80%	76.11%	-14.16%	96.93%	2.73%	48.76%	-41.51%	98.61%	4.41%	19.70%	-70.56%
2	11	1.5	0.75	98.23%	0.05%	92.77%	-5.18%	98.93%	0.75%	75.30%	-22.65%	99.36%	1.18%	42.96%	-54.99%
3	11	0.75	1.5	99.37%	0.52%	92.77%	-5.85%	99.87%	1.02%	75.30%	-23.32%	99.97%	1.12%	42.96%	-55.66%
4	12	1.5	1.5	96.37%	-0.37%	85.49%	-9.11%	98.38%	1.64%	63.59%	-31.01%	99.29%	2.55%	33.38%	-61.22%
5	12	1.5	0.75	99.23%	0.03%	96.58%	-2.52%	99.55%	0.34%	86.00%	-13.10%	99.73%	0.53%	60.50%	-38.61%
6	12	0.75	1.5	99.73%	0.24%	96.58%	-2.84%	99.95%	0.45%	86.00%	-13.43%	99.99%	0.50%	60.50%	-38.93%
7	13	1.5	1.5	98.12%	-0.16%	91.83%	-5.35%	99.19%	0.91%	76.11%	-21.08%	99.66%	1.38%	48.76%	-48.42%
8	13	1.5	0.75	99.69%	0.02%	98.51%	-1.12%	99.82%	0.15%	92.77%	-6.86%	99.90%	0.22%	75.30%	-24.34%
9	13	0.75	1.5	99.89%	0.10%	98.51%	-1.26%	99.98%	0.19%	92.77%	-7.00%	100.00%	0.21%	75.30%	-24.48%
10	14	1.5	1.5	99.09%	-0.06%	95.71%	-2.90%	99.62%	0.47%	85.49%	-13.12%	99.84%	0.70%	63.59%	-35.02%
11	14	1.5	0.75	99.88%	0.01%	99.40%	-0.46%	99.93%	0.06%	96.58%	-3.27%	99.96%	0.09%	86.00%	-13.86%
12	14	0.75	1.5	99.96%	0.04%	99.40%	-0.52%	99.99%	0.07%	96.58%	-3.33%	100.00%	0.08%	86.00%	-13.92%

[a] $\Delta = SL_G^c - SL_G^p$ [b] $\Delta = SL_Z^c - SL_Z^p$

By comparing Tables 3 and 4, and Tables 5 and 6 we notice that the stock reservation policy results into much lower service levels for Silver customers even when the response times increase. However, when the response time for Gold customers increases (compare Tables 4 and 5), the differences between the service levels for Gold customers in the two policies decrease.

Table 6 Comparison between stock reservation policy and pipeline stock priority policy, $L = 3$, $\tau = 0.5$, $\zeta = 1$.

Inputs				$K = 2$				$K = 4$				$K = 6$			
Cases	S	λ_G	λ_Z	SL_G^c	Δ^a	SL_Z^c	Δ^b	SL_G^c	Δ^a	SL_Z^c	Δ^b	SL_G^c	Δ^a	SL_Z^c	Δ^b
1	11	1.5	1.5	93.40%	-0.80%	84.72%	-7.76%	96.93%	2.73%	60.63%	-31.86%	98.61%	4.41%	28.51%	-63.98%
2	11	1.5	0.75	98.23%	0.05%	95.97%	-2.49%	98.93%	0.75%	83.11%	-15.36%	99.36%	1.18%	53.21%	-45.26%
3	11	0.75	1.5	99.37%	0.52%	95.97%	-3.17%	99.87%	1.02%	83.11%	-16.04%	99.97%	1.12%	53.21%	-45.94%
4	12	1.5	1.5	96.37%	-0.37%	91.61%	-5.12%	98.38%	1.64%	74.40%	-22.33%	99.29%	2.55%	44.57%	-52.16%
5	12	1.5	0.75	99.23%	0.03%	98.29%	-1.16%	99.55%	0.34%	91.34%	-8.11%	99.73%	0.53%	70.29%	-29.15%
6	12	0.75	1.5	99.73%	0.24%	98.29%	-1.37%	99.95%	0.45%	91.34%	-8.32%	99.99%	0.50%	70.29%	-29.37%
7	13	1.5	1.5	98.12%	-0.16%	95.74%	-2.62%	99.19%	0.91%	84.72%	-13.63%	99.66%	1.38%	60.63%	-37.73%
8	13	1.5	0.75	99.69%	0.02%	99.33%	-0.44%	99.82%	0.15%	95.97%	-3.80%	99.90%	0.22%	83.11%	-16.67%
9	13	0.75	1.5	99.89%	0.10%	99.33%	-0.54%	99.98%	0.19%	95.97%	-3.90%	100.00%	0.21%	83.11%	-16.77%
10	14	1.5	1.5	99.09%	-0.06%	97.99%	-1.22%	99.62%	0.47%	91.61%	-7.61%	99.84%	0.70%	74.40%	-24.82%
11	14	1.5	0.75	99.88%	0.01%	99.76%	-0.16%	99.93%	0.06%	98.29%	-1.62%	99.96%	0.09%	91.34%	-8.57%
12	14	0.75	1.5	99.96%	0.04%	99.76%	-0.19%	99.99%	0.07%	98.29%	-1.66%	100.00%	0.08%	91.34%	-8.61%

[a] $\Delta = SL_G^c - SL_G^p$ [b] $\Delta = SL_Z^c - SL_Z^p$

The experiments suggest that for low arrival rates or sufficient stock, the pipeline stock priority policy achieves comparable service levels for Gold customers to the stock reservation policy, while improving considerably the service level for Silver customers. For example, in Case 12, Table 3, the difference between the two policies in the service level for Gold customers is less than 0.3% whereas,

for Silver customers, the pipeline stock priority policy can improve their service level up to 20.46%. In the same time, when less stock is available, the arrival rate of Silver customers is high and a high service level is required for Gold customers while the service level for Silver customers is not very important, the stock reservation policy guarantees a better service level for Gold customers than the pipeline stock priority policy.

7. Conclusion

In this paper, we have compared two customer differentiation policies: stock reservation and pipeline stock priority for high priority customers. In particular, we have considered customer-focused performance metrics, namely the response time and the level of guarantee, because of their capability to best match customers' expectation. We provide exact analytical expressions for the service level within reponse time fot the stock reservation policy. For the pipeline stock priority policy, we provide accurate approximation methods. Via extensive numerical experiments, we found that, in cases with low arrival rate and sufficient available stock, the pipeline priority policy offers comparable service levels to Gold customers as the stock reservation policy, while offering a higher service level to Silver customers. When the arrival rate of Silver customers is relatively high and stock is scarce, the stock reservation policy insures a higher service level for Gold customers, however, at the expense of the service level for Silver customers. The numerical results indicate that it is better to use the stock reservation policy when Gold customers are much more valuable than Silver customers, and the service level offered to the latter ones is not of importance; on the other hand, when one desires to offer good service levels to both types of customers given differentiated response times, the pipeline priority policy is a better option.

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

Appendix A: Proofs of results in Section 4

Proof of Proposition 1 For calculating the distribution of the waiting time of a Silver customer, we use the equivalent (SSS) system presented in Section 4. Due to the FCFS rule for fulfilling shortfalls for Gold backorders, a Silver customer is waiting only if there are more than $S-K$ items in the pipeline. The waiting time of a Silver customer thus corresponds to the waiting time in a continuous review base stock inventory model, with basestock level $S-K$ and arrival rate λ. Based on Proposition 1 in Yang et al. (2012), it is given by

$$P(W_Z^c \leq t) = 1 - \mathbf{Erl}_{S-K,\lambda}(L-t) = \mathbf{Po}(S-K-1, \lambda(L-t)).$$

□

Proof of Proposition 2 Observe that a Gold customer is waiting only if he sees at his arrival at least S items in pipeline and will get an item that was used to restore the shortfall. In order to look at the the waiting time of Gold customers, we will look at the waiting time for each item used to restore a shortfall for Gold, and thus used to eventually serve Gold customers. Tag an item used to restore a shortfall for Gold customers at time $\tilde{t}$ when it was ordered. Consider now the second stage of the equivalent (SSS) system, described in Section 4. Due to the FCFS rule used to clear shortfalls and backorders, the second stage of the (SSS) inventory model behaves as a classical base stock inventory system with stock level $S-K$ and arrival rate λ. Hence the tagged item will be used to restore the $S-K$-th shortfall/ backorder at time $\tilde{t}+Y_{S-K,\lambda}$, where $Y_{S-K,\lambda}$ is the time of arrival of the $S-K$-th future demand at the second stage. Since the tagged item is used to restore a shortfall, it will be given to the $K-$th Gold arrival after $\tilde{t}+Y_{S-K,\lambda}$. Let W_G^c be the waiting time for the tagged item. Clearly, $W_G^c \geq t$ if and only if $Y_{S-K,\lambda}+Y_{K,\lambda_G} \leq L-t$.

Hence,

$$\begin{aligned} P(W_G^c \geq t) &= P(Y_{S-K,\lambda}+Y_{K,\lambda_G} \leq L-t), \\ &= \mathbf{Erl}_{S-K,\lambda} * \mathbf{Erl}_{K,\lambda_G}(L-t). \end{aligned}$$

The statement in the proposition now follows. □

Appendix B: Proofs of results in Section 5.2

The following notations will be used in the proofs. For $m,p,s \in \mathbf{N}$ and intervals $I_1,...,I_s$, we denote by

A_m - the time between the arrival of the Silver customer and the arrival of the m-th item in pipeline

U_p - the time between the arrival of item $n-S+p$ and item $n-S+p+1$ in pipeline.

$N_G(I_1,...,I_s)$ - a vector having as k-th component the number of Gold customers that arrive in time interval I_k, $k=1,...,s$.

$\beta=\frac{n+1}{L}$.

Before proving Lemma 1 and Lemma 2, we state a few prelimiary results that will be useful in the sequel.

Tag a Silver customer who sees n items in pipeline at arrival. If $n\geq S$ this implies that there are $n-S$ customers waiting in front of him. We take the arrival of the Silver customer as time reference. Assume that items in pipeline are numbered in the order of the remaining time in pipeline, from the moment the Silver customer arrived. Note that in a system without priorities, the tagged Silver customer would get $n-S+1$-th item in pipeline. In a system with priorities, this happens only if no Gold customer arrives between the arrival of the Silver customer and item $n-S+1$ in pipeline, in other words, if $N_G(A_{n-S+1})=0$. The Silver customer will get item $n-S+2$ in pipeline only if one Gold customer arrives before item $n-S+1$ in pipeline arrives and no Gold customer arrives between item $n-S+1$ and $n-S+2$. In other words, if $N_G(A_{n-S+1},U_1)=(1,0)$. Similarly, one can argue that the Silver customer gets item $n-S+3$ in pipeline if $N_G(A_{n-S+1},U_1,U_2)\in\{(1,1,0),(2,0,0)\}$ and item $n-S+4$ if $N_G(A_{n-S+1},U_1,U_2,U_3)\in\{(1,1,1,0),(1,2,0,0),(2,1,0,0),(2,0,1,0),(3,0,0,0)\}$. The following lemma gives the necessary and sufficient cnditions for a Silver customer to get the $n-S+k+1$- th item in pipeline, $k\geq 1$.

LEMMA 3. *A Silver customer who sees n items in pipeline at arrival, $n\geq S$, gets item $n-S+k+1$, $k\geq 1$, if and only if the following three conditions are satisfied:*

(i) $N_G(A_{n-S+1})\geq 1$

(ii) If $k\geq 2$, *for each* l, $1\leq l\leq k-1$, $N_G(A_{n-S+1})+\sum_{s=1}^{l}N_G(U_s)\geq l+1$

(iii) $N_G(A_{n-S+1})+\sum_{i=1}^{k}N_G(U_i)=k$.

Proof First, assume that the Silver customer gets item $n-S+k+1$, $k\geq 1$. We prove that in this case conditions (i)- (iii) are satisfied.

(i) If $N_G(A_{n-S+1})=0$, then the Silver customer gets item $n-S+1$, which contradicts the claim that the Silver customer gets item $n-S+k+1$, with $k\geq 1$.

(ii) Suppose that $k\geq 2$ and there is an l, $1\leq l<k$ such that $N_G(A_{n-S+1})+\sum_{s=1}^{l}N_G(U_s)\leq l$. Let $\tilde{l}$ be the smallest l with this property. This means that in the interval $A_{n-S+1}\cup U_1...\cup U_{\tilde{l}}$ at most $\tilde{l}$ Gold customers have arrived. On the other hand, in this time, item $n-S+1$,..., $n-S+1+\tilde{l}$ in pipeline have arrived in stock. Since the number of items that arrived in stock between the arrival of the Silver customer and of item

$n-S+\tilde{l}+1$ in pipeline, is larger than the number of Gold customers that have arrived in the same time interval, the Silver customer gets a pipeline item that arrived before item $n-S+\tilde{l}+1$ or item $n-S+\tilde{l}+1$. Since $\tilde{l}<k$, this contradicts the fact that the Silver customer gets item $n-S+1+k$ in pipeline.

(iii) If the Silver customer gets item $n-S+2$, $N_G(A_{n-S+1},U_1)=(1,0)$. Hence, $N_G(A_{n-S+1})+N_G(U_1)=1$. Suppose that $k\geq 2$. Based on (ii), we know that for $1\leq l<k$, in the intervals $A_{n-S+1}\cup U_1...\cup U_l$ more Gold customers than pipeline items have arrived. If $N_G(A_{n-S+1})+\sum_{i=1}^{k}N_G(U_i)>k$, then this would also be the case in $A_{n-S+1}\cup U_1...\cup U_k$, implying that the service of the Silver customer is postponed after item $n-S+1+k$ arrives in stock.

Assume now that (i)-(iii) are satisfied. Recall that in a system without priorities, the Silver customer would get item $n-S+1$ in a pipeline. If $k=1$, conditions (i) and (ii) imply that $N_G(A_{n-s+1},U_1)=(1,0)$, so the Silver customer will get item $n-S+2$.

Assume $k\geq 2$. Based on condition (i), we can conclude that the Silver customer will not get item $n-S+1$ in pipeline. Based on (ii) we know that in the interval $A_{n-S+1}\cup U_1...\cup U_l$, $l\leq k-1$ more Gold customers than pipeline items have arrived. Hence, the Silver customer will get a pipeline item that arrives after item $n-S+k$. Moreover, (iii) implies that in $A_{n-S+1}\cup U_1...\cup U_k$ exactly k Gold customers have arrived. Thus, $N_G(U_k)=0$ and by the time item $n-S+k+1$ joins the stock, all the waiting Gold customers have been served. The Silver customer gets item $n-S+k+1$ in pipeline. □

For $k\geq 1$, define the sets V_k by $V_1=\{(1,0)\}$ and

$$V_k=\{(n_0,...,n_{k-1},n_k)|\sum_{i=0}^{l}n_i\geq l+1,\ \text{for } 1\leq l\leq k-1,\sum_{i=0}^{k-1}n_i=k \text{ and } n_k=0\}.$$

Note that (ii) and (iii) also imply that if a Silver customer who at arrival sees n items in pipeline, gets item $n-S+k+1$, $N_G(U_k)=0$.

REMARK 1. Lemma 3 implies that a Silver customer who sees $n\geq S$ items in pipeline at arrival, gets item $n-S+k+1$, $k\geq 1$, if and only if $N_G(A_{n-S+1},U_1,...,U_k)\in V_k$.

Following Brualdi (2004) Chapter 8, the set V_k can be given the following geometric interpretation. Call a *lattice path* a path p in the Euclidian plane that contains only horizontal segments from left to right and upwards vertical segments. A lattice path p is called a *subdiagonal lattice path* if it lies under or touches the diagonal $y=x$. The set V_k can be interpreted as the set of lattice paths from $(0,0)$ to $(k,k+1)$, formed by a subdiagonal lattice path from $(0,0)$ to (k,k) and a vertical segement from (k,k) to $(k,k+1)$. The subdiagonal lattice path has at $y=l$, $0\leq l\leq k-1$ an horizontal segment of length n_l.

Next we introduce two other sets that will be useful in further analysis. Recall that if a Silver customer sees at arrival $n \geq S$ items in pipeline, he will get item $n-S+1+k$, with $k \geq 1$. For $k \leq S$ he will get an item he sees in pipeline upon arrival or the item he orederes (if $k = S$), while for $k > S$, will get an item ordered after his arrival. Hence, by Assumption 3, $A_{n-S+1} \sim \mathbf{Erl}_{n-S+1,\beta}(\cdot)$ and $U_k \sim \mathbf{Exp}_{\beta}(\cdot)$, for $1 \leq k \leq S$. Moreover, since items $n+2, ...n-S+1+k$ are ordered after the Silver customer has arrived, $U_k \sim \mathbf{Exp}_{\lambda}$ for $k > S$. Due to the fact that $\{A_{n-S+1}\}$, $\{U_1, ..., U_S\}$ and $\{U_{S+1}..., U_{n-S+1+k}\}$, $k > S$ follow different distributions, in our analysis, we will make distinction between the number of Gold customers that arrive during intervals in each set. Together with the sets V_k, the following sets will be used in the proofs:

$$V_k(a) = \{(n_1, ..., n_k) | (a, n_1, ..., n_k) \in V_k\} \tag{6}$$

and

$$V_k(a, B) = \{(n_1, ..., n_k) | (a, n_1, ..., n_k) \in V_k, \sum_{i=1}^{S} n_i = B\}. \tag{7}$$

Similar to the interpretation of V_k in Remark 1, we observe that if a Silver customers sees $n \geq S$ items in pipeline at arrival and a Gold customers arrive during A_{n-S+1}, he will get item $n-S+1+k$ if and only if $N_G(U_1, ..., U_k) \in V_k(a)$. Geometrically, $V_k(a)$ can be interpreted as the set of lattice paths from $(0,0)$ to $(k, k+1)$, with an horizontal segment of length a at $y = 0$, a subdiagonal lattice path from $(a, 1)$ to (k, k) and a vertical segment from (k, k) to $(k, k+1)$. A similar interpretation can be given to $V_k(a, B)$.

Denote by $SP((a,b) : (c,d))$ the set of subdiagonal lattice paths p from (a,b) to (c,d).

Next lemma gives a characterization of $SP((a,b) : (c,d))$, V_k, $V_k(a)$ and $V_k(a,B)$ that will prove useful in further analysis. Part of the results ((a) and (b)) are taken from Brualdi (2004).

LEMMA 4. *The sets V_k satisfy the following properties:*

(a) (Brualdi (2004)) $|V_k| = |SP((0,0) : (k,k))| = \frac{1}{k+1}\binom{2k}{k}$.

(b) $|SP((0,0) : (p,q))| = \frac{p-q+1}{p+1}\binom{p+q}{q}$.

(c) For $0 < b \leq a \leq k$, $|SP((a,b),(k,k))| = \frac{a+1-b}{k-b+1}\binom{2k-a-b}{k-a}$.

(d) For $a \geq 1$, $|V_k(a)| = \frac{a}{k}\binom{2k-a-1}{k-1}$.

(e) For $a \geq b$, $p \geq q$, $p \geq a$ *and* $q \geq b$,

$$|SP((a,b) : (p,q)) = \begin{cases} \binom{p-a+q-b}{p-a} - \binom{p-a+q-b}{p-b+1}, a < q \\ \binom{p-a+q-b}{p-a}, \quad \textit{otherwise.} \end{cases}$$

(f) For $k \geq S$ *and* $B \geq 0$,

$$|V_k(a,B)| = \begin{cases} \frac{a+B-S}{k-S+1}\binom{2k-a-B-S}{k-S-1}\left(\binom{B+S-1}{B} - \binom{B+S-1}{a+B}\right) \textit{ for } a < S \\ \frac{a+B+1-S}{k-S}\binom{2k-a-B-S-1}{k-S-1}\binom{B+S-1}{B}, \textit{ for } a \geq S. \end{cases}$$

Proof (a), (b) See Brualdi (2004), Chapter 8, Theorem 8.5.2 and 8.5.3. Note that $\frac{1}{k+1}\binom{2k}{k}$ is equal to the well known k-th Catalan number.

(c) It is easy to see that for $b \leq a \leq k$ the number of subdiagonal lattice paths from (a,b) to (k,k) is the same as the number of subdiagonal lattice paths from $(0,0)$ to $(k-b,k-a)$. Based on (b), this is equal to $\frac{a-b+1}{k-b+1}\binom{2k-a-b}{k-a}$.

(d) Based on the geometric interpretation of $V_k(a)$, we can conclude that $|V_k(a)|$ is equal to the number of subdiagonal lattice paths from $(a,1)$ to (k,k). Hence, based on (c), $|V_k(a)| = \frac{a}{k}\binom{2k-a-1}{k-a}$.

(e) In Brualdi (2004) Chapte 8, Theorem 8.5.1, is shown that for $p \geq a$ and $q \geq b$, the number of lattice paths from (a,b) to (p,q) is equal to $\binom{p-a+q-b}{p-a}$. Note that these paths may cross the diagonal.

If $q \leq a$, all paths from (a,b) to (p,q) are subdiagonal. Next consider the case $q > a$.

Denote by $DP((a,b),(p,q))$ the set of lattice paths from (a,b) to (p,q) that cross the diagonal. By the definition of a subdiagonal path,

$$|SP((a,b),(p,q))| = \binom{p-a+q-b}{p-a} - |DP((a,b),(p,q))|. \tag{8}$$

Observe that the number of paths in $DP((a,b),(p,q))$ coincides with the number of lattice paths from $(a,b-1)$ to $(p,q-1)$ that touch or cross the diagonal $y=x$. This correspondence can be seen by shifting the paths in $DP((a,b),(p,q))$ one unit down. Consider now a lattice path from $(a,b-1)$ to $(p,q-1)$ that touches or crosses the diagonal at (d,d). Let γ_1 be the path from $(a,b-1)$ to (d,d) and γ_2 the path from (d,d) to $(p,q-1)$. Reflect γ_1 on the diagonal and call the resulting path γ_1^*. γ_1^* is a path from $(b-1,a)$ to (d,d). The path γ_1^*, continued with γ_2 is a lattice path from $(b-1,a)$ to $(p,q-1)$. Observe that since $b-1<a$ and $p>q-1$, all the lattice paths from $(b-1,a)$ to $(p,q-1)$ cross the diagonal. In a similar way, we can associate to every lattice path between $(b-1,a)$ and $(p,q-1)$, a path from $(a,b-1)$ to $(p,q-1)$ that crosses or touches the diagonal. Hence, $|DP((a,b),(p,q))|$ equals the number of lattice paths from $(b-1,a)$ to $(p,q-1)$. More precisely,

$$|DP((a,b),(p,q))| = \binom{p-a+q-b}{p-b+1}. \tag{9}$$

From (8) and (9) follows that

$$|SP((a,b),(p,q))| = \binom{p-a+q-b}{p-a} - \binom{p-a+q-b}{p-b+1}.$$

(f) To each element of $V_k(a,B)$, one can associate a subdiagonal lattice path consisting from two subpaths:

one from $(a,1)$ to $(a+B,S)$ and one from $(a+B,S+1)$ to (k,k). Observe also that the definition of $V_k(a,B)$ implies that $a+B \geq S+1$. Hence,

$$|V_k(a,B)| = |SP((a,1):(a+B,S)| \cdot |SP((a+B,S+1):(k,k))|.$$

Combining e), and c) we obtain

$$|V_k(a,B)| = \begin{cases} \frac{a+B-S}{k-S}\binom{2k-a-B-S-1}{k-S-1}\left(\binom{B+S-1}{B} - \binom{B+S-1}{a+B}\right) & \text{for } a < S \\ \frac{a+B-S}{k-S}\binom{2k-a-B-S-1}{k-S-1}\binom{B+S-1}{B}, & \text{for } a \geq S. \end{cases}$$

□

Next Lemma gives the probability distribution of the number of Gold customers that delay the service of the Silver customer.

LEMMA 5. *The distribution of the number of Gold customers arriving during a time T that is* ***Erl****(m,β) distributed, is given by*

$$P(N_G(T)=a) = \binom{a+m-1}{a}\left(\frac{\lambda_G}{\lambda_G+\beta}\right)^a\left(\frac{\beta}{\lambda_G+\beta}\right)^m.$$

Proof By conditioning on the variable T, we get

$$\begin{aligned} P(N_G(T)=a) &= \int_0^\infty e^{-\lambda_G t}\frac{(\lambda_G t)^a}{a!}e^{-\beta t}\frac{\beta(\beta t)^{m-1}}{(m-1)!}dt \\ &= \binom{a+m-1}{a}\left(\frac{\lambda_G}{\lambda_G+\beta}\right)^a\left(\frac{\beta}{\lambda_G+\beta}\right)^m, \end{aligned} \tag{10}$$

where for the last equality we have used that $\int_0^\infty e^{-(\lambda_G+\beta)t}\frac{(\lambda_G+\beta)^{a+m}t^{a+m-1}}{(a+m-1)!}dt = 1$, as the function under the integral is the density function of an $\mathbf{Erl}(a+m,\lambda_G+\beta)$ variable. □

From Lemma 5 and Assumption 3 which implies that $A_{n-S+1} \sim \mathbf{Erl}_{n-S+1,\beta}(\cdot)$, $U_m \sim \mathbf{Exp}_\beta(\cdot)$ for $1 \leq m \leq S$ and $U_m \sim \mathbf{Exp}_\lambda(\cdot)$ for $m > S$, we obtain the following Corrolary.

COROLLARY 1. *If a Silver customer sees at arrival n items in pipeline, $n \geq S$, then,*

(a) $P(N_G(A_{n-S+1})=a) = \binom{n-S+a}{a}\left(\frac{\lambda_G}{\lambda_G+\beta}\right)^a\left(\frac{\beta}{\lambda_G+\beta}\right)^{n-S+1}$.

(b) For $1 \leq m \leq S$, $P(N_G(U_m)=a) = \frac{\beta}{\lambda_G+\beta}\left(\frac{\lambda_G}{\lambda_G+\beta}\right)^a$.

(c) For $m \geq S+1$, $P(N_G(U_m)=a) = \frac{\lambda}{\lambda_G+\lambda}\left(\frac{\lambda_G}{\lambda_G+\lambda}\right)^a$.

Proof of Lemma 1 The Silver customer gets item $n-S+1$ if no Gold customers arrive during A_{n-S+1}. By Corrolary 1, this is equal to

$$P(N_G(A_{n-S+1})=0) = \left(\frac{\beta}{\lambda_G+\beta}\right)^{n-S+1}.$$

If $1 \leq k \leq S$, the Silver customer gets an item he sees in pipeline at his arrival or the item ordered at his arrival. As shown in Lemma 3, the Silver customer gets item $n-S+k+1$ if and only if $N_G(A_{n-S+1}, U_1, ..., U_{k-1}, U_k) \in V_k$. Since the intervals A_{n-S+1},U_1,..., U_k are independent by Assumption 3, the probability $p_{n,k}$ can now be calculated as follows:

$$\begin{aligned} p_{n,k} &= P(N_G(A_{n-S+1}, U_1, ..., U_{k-1}, U_k) \in V_k) \\ &= \sum_{a=1}^{k} P(N_G(A_{n-S+1}) = a) P(N_G(U_1, ..., U_{k-1}, U_k) \in V_k(a)), \end{aligned} \tag{11}$$

where for $a \geq 1$, $V_k(a) = \{(a, ..., n_k) | (a, ..., n_k) \in V_k\}$. Based on Lemma 4 d), $|V_k(a)| = \frac{a}{k}\binom{2k-a-1}{k-a}$.

From Corollary 1 now follows

$$P(N_G(A_{n-S+1}) = a) = \binom{n-S+a}{a} \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^a \left(\frac{\beta}{\lambda_G+\beta}\right)^{n-S+1}. \tag{12}$$

Moreover,

$$\begin{aligned} P(N_G(U_1, ..., U_k) \in V_k(a)) &= \sum_{(n_1,...,n_k)\in V_k(a)} P(N_G(U_1) = n_1)....P(N_G(U_{k-1}) = n_k) P(N_G(U_k) = 0) \\ &= \sum_{(n_1,...,n_k)\in V_k(a)} \left(\frac{\beta}{\lambda_G+\beta}\right)^k \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{\sum_{i=1}^k n_i} \\ &= \left(\frac{\beta}{\lambda_G+\beta}\right)^k \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{k-a} |V_k(a)| \\ &= \left(\frac{\beta}{\lambda_G+\beta}\right)^k \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{k-a} \frac{a}{k}\binom{2k-a-1}{k-a}. \end{aligned} \tag{13}$$

In the second equality above we have used the fact that if $(n_1, ..., n_k) \in V_k(a)$, then $(a, n_1, ..., n_k) \in V_k$, which means that $a + \sum_{i=1}^k n_i = k$. Finally, combining (11)-(13) we obtain

$$p_{n,k} = \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^k \left(\frac{\beta}{\lambda_G+\beta}\right)^{n-S+k+1} \sum_{a=1}^{k} \frac{a}{k}\binom{n-S+a}{a}\binom{2k-a-1}{k-a}.$$

□

Proof of Lemma 2 Let $\beta = \frac{n+1}{L}$. As shown in Lemma 3 , the Silver customer gets item $n-S+k+1$ if and only if $N_G(A_{n-S+1}, U_1, ..., U_{k-1}, U_k) \in V_k$. By the independence of $A_{n-S+1}, U_1, ..., U_S..., U_k$ we obtain

$$\begin{aligned} p_{n,k} &= P(N_G(A_{n-S+1}, U_1, ..., U_k) \in V_k) \\ &= \sum_{a=1}^{k} P(N_G(A_{n-S+1}) = a) \sum_{(n_1,...,n_k)\in V_k(a)} \prod_{m=1}^{k} P(N_G(U_m) = n_m), \end{aligned} \tag{14}$$

where for $a \geq 1$, $V_k(a)$ is defined by (6). Recall that for $m \geq S+1$, $U_m \sim \mathbf{Exp}_\lambda(\cdot)$. Also, $n_k = 0$ for $(a, n_1, ..., n_k) \in V_k$. Based on Corrolary 1 (c), we now obtain

$$\prod_{m=S+1}^{k} P(N_G(U_m) = n_m) = \prod_{m=S+1}^{k} \frac{\lambda}{\lambda_G+\lambda} \left(\frac{\lambda_G}{\lambda_G+\lambda}\right)^{n_m}$$

$$= \left(\frac{\lambda}{\lambda_G+\lambda}\right)^{k-S} \left(\frac{\lambda_G}{\lambda_G+\lambda}\right)^{\sum_{i=S}^{k} n_i}$$
$$= \left(\frac{\lambda}{\lambda_G+\lambda}\right)^{k-S} \left(\frac{\lambda_G}{\lambda_G+\lambda}\right)^{k-a-\sum_{m=1}^{S} n_m}, \quad (15)$$

where for the last equality we have used that $a+\sum_{m=1}^{k} n_m = k$, for $(a, n_1, ..., n_k) \in V_k$.

By Assumption 3, for $1 \le m \le S$, U_m are exponential with rate β. We conclude that

$$\prod_{m=1}^{S} P(N_G(U_m) = n_m) = \left(\frac{\beta}{\lambda_G+\beta}\right)^{S} \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{\sum_{m=1}^{S} n_m}. \quad (16)$$

Combining (15) and (16), we obtain that for $(n_1, ..., n_m) \in V_k(a)$,

$$\prod_{m=1}^{k} P(N_G(U_m) = n_m) = \left(\frac{\lambda}{\lambda_G+\lambda}\right)^{k-S} \left(\frac{\beta}{\lambda_G+\beta}\right)^{S} \left(\frac{\lambda_G}{\lambda_G+\lambda}\right)^{k-a-\sum_{m=1}^{S} n_m} \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{\sum_{m=1}^{S} n_m}. \quad (17)$$

Note that for $(n_1, ..., n_k) \in V_k(a)$, $a+\sum_{i=1}^{S} n_i \ge S+1$. For B such that $\max\{S+1-a, 0\} \le B \le k-a$, let $V_k(a, B)$ be defined as in (7).

Equation (17) implies

$$\sum_{(n_1,...,n_k)\in V_k(a)} \prod_{m=1}^{k} P(N_G(U_m) = n_m) =$$
$$\left(\frac{\lambda}{\lambda_G+\lambda}\right)^{k-S} \left(\frac{\beta}{\lambda_G+\beta}\right)^{S} \sum_{B=\max\{S+1-a,0\}}^{k-a} \left(\frac{\lambda_G}{\lambda_G+\lambda}\right)^{k-a-B} \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{B} |V_k(a, B)| \quad (18)$$

By Lemma 4 f),

$$|V_k(a, B)| = \begin{cases} \frac{a+B-S}{k-S}\binom{2k-a-B-S-1}{k-S-1}\left(\binom{B+S-1}{B} - \binom{B+S-1}{a+B}\right) & \text{for } a < S \\ \frac{a+B-S}{k-S}\binom{2k-a-B-S-1}{k-S-1}\binom{B+S-1}{B}, & \text{for } a \ge S. \end{cases} \quad (19)$$

Combining (14) and (18) we obtain

$$p_{n,k} = \sum_{a=1}^{k} \binom{n-S+a}{a} \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{a} \left(\frac{\beta}{\lambda_G+\beta}\right)^{n-S+1} \left(\frac{\lambda}{\lambda_G+\lambda}\right)^{k-S} \left(\frac{\beta}{\lambda_G+\beta}\right)^{S} \cdot$$
$$\sum_{B=\max\{S+1-a,0\}}^{k-a} \left(\frac{\lambda_G}{\lambda_G+\lambda}\right)^{k-a-B} \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{B} |V_k(a, B)| =$$
$$= \left(\frac{\lambda}{\lambda_G+\lambda}\right)^{k-S} \left(\frac{\beta}{\lambda_G+\beta}\right)^{n+1} \sum_{a=1}^{k} \binom{n-S+a}{a} \sum_{B=\max\{S+1-a,0\}}^{k-a} \left(\frac{\lambda_G}{\lambda_G+\lambda}\right)^{k-B} \left(\frac{\lambda_G}{\lambda_G+\beta}\right)^{B} q_k(a, B),$$

with $q_k(a, B) = |V_k(a, B)|$ given by (19). □

References

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

Alvarez, E. M., M. C. Van der Heijden, W. H. M. Zijm. 2012. The selective use of emergency shipments for service-contract differentiation. *International Journal of Production Economics* **143**(2) 518–526.

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

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

Brumelle, S.L. 1978. A generalization of erlang's loss system to state dependent arrival and service rates. *Mathematics of Operations Research* **3**(1) 10–16.

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

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

Fritzsche, R., R. Lasch. 2012. An integrated logistics model of spare parts maintenance planning within the aviation industry. *World Academy of Science, Engineering and Technology* **68**.

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.

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.

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.

Yang, G., R. Dekker, A. F. Gabor, S. Axsäter. 2012. Service parts inventory control with lateral transshipment and pipeline stock flexibility. *International Journal of Production Economics* **142**(2) 278–289.

ERIM Report Series *Research in Management*	
ERIM Report Series reference number	ERS–2014–003–LIS
Date of publication	2014–02–14
Version	14–02–2014
Number of pages	31
Persistent URL for paper	http://hdl.handle.net/1765/50502
Email address corresponding author	gabor@ese.eur.nl

Address	Erasmus Research Institute of Management (ERIM) RSM Erasmus University / Erasmus School of Economics Erasmus University Rotterdam PO Box 1738 3000 DR Rotterdam, The Netherlands Phone: +31104081182 Fax: +31104089640 Email: info@erim.eur.nl Internet: http://www.erim.eur.nl

Availability	The ERIM Report Series is distributed through the following platforms: RePub, the EUR institutional repository Social Science Research Network (SSRN) Research Papers in Economics (RePEc)
Classifications	The electronic versions of the papers in the ERIM Report Series contain bibliographic metadata from the following classification systems: Library of Congress Classification (LCC) Journal of Economic Literature (JEL) ACM Computing Classification System Inspec Classification Scheme (ICS)