European Journal of Operational Research 000 (2018) 1–9

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing and Logistics

A base-stock inventory model with service differentiation and response time guarantees

Adriana F. Gabor[a,*], Lars A. van Vianen[b], Guangyuan Yang[b], Sven Axsäter[c]

[a] College of Business and Economics, United Arab Emirates University, Al Ain, Abu-Dhabi, United Arab Emirates
[b] Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam 3062 PA, Netherlands
[c] Department of Industrial Management and Logistics, Lund University, Lund S-221 00, Sweden

ARTICLE INFO

Article history:
Received 15 June 2017
Accepted 15 February 2018
Available online xxx

Keywords:
Inventory
Service differentiation
Response time
Queueing

ABSTRACT

In response to customer specific service time guarantee requirements, service providers can offer differentiated services. Although response time agreements offer more flexibility, most of the theoretical models for customer differentiation are based on fill rates, due to the difficulty of calculating the response time distribution in the presence of priorities. 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. For this model, we derive the exact distribution of the response time (within lead time) for the lower priority class and provide a new closed-form approximation for the fill rate for high priority customers. Via numerical experiments, we show that the fill rate approximation performs comparable to the best approximations in the literature, while its implementation requires only elementary operations.

1. Introduction

Offering different service levels to different categories of customers arises frequently in practice. For example, a provider of IT services will be required to react immediately when a mainframe computer in a stock exchange fails, whereas a short delay in response when the mainframe of a school library fails might be accepted. A second example is that of a spare parts service provider who has two categories of customers: some keep inventory on-site and some do not. Usually, when a part fails at a customer who keeps on-site stock, the part is replaced from the on-site stock and an order is placed with the provider. For this type of customers, an immediate response is less critical. On the other hand, for the customers who do not keep inventory on-site, an immediate response might be requested. A third example is that of a retailer who sells items online and through a brick-and-mortar store (an omni-channel retailer), and who responds faster to the customers who visit the store.

The challenge of inventory managers with differentiated customers is to offer the agreed service level while having a minimal capital investment in inventory. In order to take advantage of the economies of scale obtained by pooling inventory while delivering differentiated services, researchers proposed to use critical level policies. These policies reserve a part of inventory for high priority customers and pool the rest of the resources (Arslan, Graves, & Roemer, 2007; Dekker, Kleijn, & De Rooij, 1998; Deshpande, Cohen, & Donohue, 2003; Nahmias & Demmy, 1981; Veinott, 1965; Vicil & Jackson, 2016).

Most of the inventory literature with differentiated customers focuses on minimizing expected on-hand inventory, while imposing a desired level on the fill rate. The fill rate, however, is not always the most appropriate measure for the offered service level. 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 instead of fillrates. 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, Alvarez, & Schutten, 2012). In an omni-channel environment, a response time guarantee seems more appropriate than the fill rate in case of online customers, who are usually willing to wait for service. Despite their applicability, imposing response time constraints is not straightforward, due to the difficulty of deriving response time distributions in inventory systems with priorities.

In this paper we focus on incorporating response time constraints in a continuous-review $(S-1,S,K)$ inventory model with two demand classes (Gold and Silver). We assume that each demand type follows a stationary Poisson process and a constant lead time. Low priority (Silver) customers are served only when the on

* Corresponding author.
E-mail addresses: adriana.gabor@uaeu.ac.ae (A.F. Gabor), lars29@live.nl (L.A. van Vianen), gyang@ese.eur.nl (G. Yang), sven.axsater@iml.lth.se (S. Axsäter).

https://doi.org/10.1016/j.ejor.2018.02.039
0377-2217/

hand inventory is greater than K and unsatisfied demand is backordered. Replenishments are first used to satisfy Gold backorders, then to refill the on hand inventory up to level K, and finally to satisfy backorders of Silver customers. Such a policy is appropriate for differentiated spare parts services, characterized by low demand and for items with high holding and shortage costs relative to the ordering costs (Alfredsson & Verrijdt, 1999; Dekker et al., 1998). Approximations for the fill rates 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, response time guarantees have not been explored in this context. Our contribution can be summarized as follows:

(i) We present an exact derivation of the distribution of the response time (within lead time) 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 main obstacle is the fact that the waiting time of a Silver customer depends not only on the number of waiting customers she sees upon arrival, but also on the number of Gold customers that arrive while she is waiting. We overcome this difficulty by using elementary lattice paths counting, a technique that has been used in the field of queuing theory by, among others, Takács (1967) and 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 common in the study of queues with priorities.
(ii) We propose a closed-form approximation for the fill rate for Gold customers, that has a comparable performance to the best approaches in the literature (Fadıloğlu & Bulut, 2010; Vicil & Jackson, 2016), while being straightforward to implement.
(iii) Via numerical experiments, we analyze the stock reduction that can be obtained by incorporating response time constraints for low priority customers in an $(S-1,S,K)$ inventory model.

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. We discuss an approximations for the fill rate for Gold customers in Section 5. An algorithm for deciding stock levels based on response time constraints for low priority customers and fill rates for Gold customers is described in Section 6. In Section 7, we validate our approximation method for the fill rate for high priority customers via extensive numerical experiments and we discuss the impact of incorporating response time constraints on the optimal base stock levels. Conclusions and further research 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 were first proposed by Veinott (1965). Topkis (1968) analyzed this policy for a periodic system with zero lead time 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 proved that within a review interval, there exist optimal, nonnegative, rationing levels for each demand class.

Ha (1997a,b) considered 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 showed that a base stock policy for the production decision and a dynamic rationing policy for inventory are optimal.

Nahmias and Demmy (1981) derived the first approximations for the expected backorders and fillrates in a continuous review (Q,R,K) inventory system with two demand classes modeled by Poisson processes 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.

Dekker et al. (1998) proposed approximations for the fill rates in an $(S-1,S,K)$ model with deterministic lead times and Poisson demand. They explored several ways of allocating the incoming replenishment items in case of stock out and concluded that the allocation method impacts significantly the duration of the stock out for the lower priority class, while having little impact on the fillrates. This in turn implies that the allocation method will impact the response time to customers.

Deshpande et al. (2003) considered the continuous review (Q,R,K) inventory system with two customer classes discussed in Nahmias and Demmy (1981). They proposed to approximate the system parameters for the priority clearing policy with the optimal parameters in a *threshold clearing* mechanism that is easier to analyse and gives results close to the optimal priority clearing policy. Deshpande and Cohen (2005) extended the analysis of this model to multiple customer classes.

Arslan et al. (2007) showed that for the model discussed by Nahmias and Demmy (1981) and Deshpande et al. (2003), the threshold clearing policy is equivalent to a policy in which backorders of the higher priority customer class, orders to replenish the stock reserved for higher priority demand and backorders of lower priority customers are served according to a FCFS discipline. They showed also that the inventory system using this policy can be analyzed by mapping it to a serial inventory system and proposed an efficient heuristic to find the policy parameters.

Vicil and Jackson (2016) analyzed the $(S-1,S,K)$ inventory system with two demand classes, under the priority clearing policy. For exponential lead times and Poisson demand, they proposed a recursive method for finding the steady state distribution of the on hand inventory and of the number of backorders of each class. They showed that the same balance equations hold in case of general lead times, 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. Under the same independence condition, Fadıloğlu and Bulut (2010) approximated the steady state probabilities for the case of constant lead times by using an embedded Markov chain. According to the numerical study in Vicil and Jackson (2016), the procedures described in Vicil and Jackson (2016) and Fadıloğlu and Bulut (2010) give the most accurate fill rates in the literature.

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 time for the lower priority customer class and on the impact of response times on stock levels. In the same time, we propose a novel close-form

Please cite this article as: A.F. Gabor et al., A base-stock inventory model with service differentiation and response time guarantees, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.02.039