European Journal of Operational Research 000 (2018) 1–10

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Production, Manufacturing and Logistics

Optimal allocation of spares to maximize the window fill rate in a two-echelon exchangeable-item repair system

Michael Dreyfuss, Yahel Giat[*]

Department of Industrial Engineering, Jerusalem College of Technology, Jerusalem, Israel

ARTICLE INFO

Article history:
Received 28 March 2017
Accepted 8 May 2018
Available online xxx

Keywords:
Inventory
Logistics
Spares allocation problem
Tolerable wait

ABSTRACT

We solve the spares allocation problem in a two-echelon, exchangeable-item repair system in which the lower echelon comprises multiple locations and the higher echelon is a single depot. We assume that customers tolerate a certain wait and therefore the optimization criterion is the window fill rate, i.e., the expected portion of customers who are served within the tolerable wait. We develop two algorithms to solve this problem. The first algorithm (FTEA) is formula-based and is suboptimal. The second algorithm (HTEA) combines simulations into the first algorithm and obtains a higher degree of accuracy at the cost of extra running time. We characterize the near-optimal solution by its degree of pooling and concentration. Pooling happens when spares are allocated to the depot and are therefore shared by all the lower-echelon locations. Concentration takes place when spares are allocated to only a few lower-echelon locations whereas the other lower-echelon locations receive no spares. We use numerical examples to compare the algorithms and to illustrate how the budget, shipment time, local repair and customer patience affect the optimal solution and degree of pooling in varying ways. Using the numerical results, we propose a third algorithm (ETEA) that obtains HTEA's output in 30% of the time.

1. Introduction

Exchangeable-item repair systems are particularly useful in technically advanced industries such as aircraft and automobile manufacturing, in which parts are expensive and scarce and are therefore repaired and reused. Frequently, these inventory systems are designed as layered systems in which the repair facilities at the lower-echelon can repair a limited range of failures whereas the higher-echelon repair facilities are more advanced and can repair all types of failures. We consider a two-echelon system that comprises multiple locations at the lower echelon and a central depot at the higher echelon. Customers arrive with failed items to any one of the system's locations and receive an operable item in exchange. The item is repaired on site if its type of failure can be repaired by the location's facilities. Otherwise, it is shipped to the central depot for repair and then returned to its original location.

To improve the system's performance, spares are placed in the different locations. Lateral transshipment is not allowed, and therefore, spares in each of the lower-echelon locations serve only its customers. In contrast, the spares at the depot serve its customers as well as the lower-echelon customers through the repair requests from the lower echelon. Thus, allotting spares to the depot is synonymous to pooling spares. Furthermore, if the total number of spares is constrained, then the degree of pooling may vary from no pooling (no spares at the depot) to full pooling (all the spares are at the depot). Each spares allocation, therefore, is characterized by its degree of pooling (i.e., the number of spares at the depot) and the lower-echelon allocation.

The spares allocation problem is defined as choosing how many spares to place in each location with the goal of maximizing the system's performance. As an alternative, we consider the optimal budget problem, which is defined as choosing the optimal number of spares and allocation to meet the system's required performance level. In this paper, the performance measure is a generalization of the fill rate. The fill rate assumes that customers penalize the firm *if* they wait. In many cases of practical interest, however, customers will tolerate a certain period of wait and therefore the firm does not incur reputation costs if the customer waits less than the tolerable wait. Accordingly, the performance measure we use is the *window fill rate*, which is defined as the probability that the customer is served *within* the tolerable wait.

The contributions of this paper are threefold. First, we extend the single-echelon (Dreyfuss & Giat, 2017a) model and develop approximation formulas for the window fill rate in a two-echelon setting. This setting is of particular importance since it allows us to investigate whether pooling spares is beneficial or detrimental to the system's performance.

[*] Corresponding author.
E-mail addresses: dreyfuss@jct.ac.il (M. Dreyfuss), yahel@jct.ac.il (Y. Giat).

https://doi.org/10.1016/j.ejor.2018.05.011
0377-2217/© 2018 Elsevier B.V. All rights reserved.

Second, we develop algorithms that solve the spares allocation problem. We first develop FTEA, a formula-based algorithm with polynomial complexity that optimizes the system's window fill rate. For each level of pooling, (i.e., for each number of spares at the depot), we determine the near-optimal lower-echelon allocation using the approximation formulas. While this algorithm is fast, the approximation error may lead to suboptimal solutions. To mitigate the approximation error, we develop HTEA in which we simulate the window fill rate for each level of pooling and then choose the optimal solution. Our use of simulation is parsimonious in that for each level of pooling we simulate *only* its corresponding near-optimal lower-echelon allocation. This prudence enables us to reduce the overall error without sacrificing considerable computation time. We show that our algorithm is applicable for systems with multiple locations, since the number of simulations needed to reach near-optimality is linear with the number of spares in the system and is independent of the number of the locations.

In practical settings, warehouses may manage numerous different item types and therefore, to further reduce running time we develop ETEA, which reduces HTEA's running time by 30% . For very large warehouses, however, even ETEA's computing time is significant. To overcome this, we use the fact that ETEA can be solved independently (and therefore, simultaneously) for each item-type. Thus, managers may acquire the services of cloud computing with multiple CPUs to determine the desired inventory levels of the warehouse in a very short time. When such services are not available, then FTEA should be used to determine spare levels for the inexpensive item-types, which are usually the majority of the items. For the fewer item-types that are particularly expensive, the slower ETEA should be used because of the considerable savings that it attains due to its higher accuracy.

Third, we complement the theory with a numerical illustration in which we compare the algorithms and investigate how the model parameters affect the optimal solution. For the optimal budget problem, FTEA may result with excessive budget and may fail to meet the target window fill rate. The FTEA's inaccuracy is more pronounced as the target fill rate is higher. For the spares allocation problem we describe how the optimal solution is affected by the target fill rate, probability to repair on site, the shipment time, the tolerable wait and the arrival rates. We show that the window fill rate changes considerably and variably with the degree of pooling. These relationships dictate varying optimal pooling levels and matching lower-echelon allocations.

The practical implication of our analysis is that pooling improves the window fill rate only when there are sufficiently many spares in the system or when customers' patience is high. Conversely, when spares are scarce or customers' tolerable wait is very small then pooling spares will decrease the system's window fill rate. Furthermore, even when it is beneficial to pool spares, the optimal degree of pooling decreases with shipment time to the depot and with the probability to repair on site.

2. Literature review

This paper contributes to the research of multi-echelon, exchangeable-item repair systems originated by Sherbrooke (1968)'s METRIC model. This body of research is presented in books such as Sherbrooke (2004) and Muckstadt (2005) and reviewed in Kennedy, Patterson, and Fredendall (2002), Wong, van Houtum, Cattrysse, and van Oudheusden (2006) and Basten and van Houtum (2014). The standard METRIC assumptions include: first come first serve service (FCFS) policy, ample repair servers, items fail according to a compound Poisson process, no lateral transshipment and an $(S-1,S)$ ("one-for-one") continuous review inventory policy.

Many METRIC-related papers (e.g., Basten, van der Heijden, & Schutten, 2012, Basten & van Houtum 2013, Ghaddar, Sakr, & Asiedu 2016, Cohen, Cohen, & Landau 2016, Dreyfuss & Giat 2017b) focus on optimizing back-order associated costs. Other papers use the fill rate as the system's performance criterion (e.g., Shtub & Simon, 1994, Caggiano, Jackson, Muckstadt, & Rappold 2007, Lien, Iravani, & Smilowitz 2014). For example, Lien et al. (2014) maximize the expected *minimum* fill rate. Caggiano et al. (2007) optimize a distribution system according to the channel fill rate, i.e., the probability being served within the lead time between the echelons, where the lead time is constant. Our system, in contrast, is a repairable-item system with random repair time and random lead time between the echelons.

The disadvantage of the fill rate is that it ignores reports that customers will tolerate a certain period of wait, (Durrande-Moreau, 1999; see also Katz, Larson, & Larson, 1991 who use the term "reasonable duration"). We incorporate this by optimizing the *window* fill rate, i.e., the probability of a random customer being served within a certain time window. Viewed as a function of the tolerable wait, the window fill rate is the waiting time distribution function and was first developed by Higa, Feyerherm, and Machado (1975) in a single echelon, $(S-1,S)$ inventory policy. They assume that the supply lead times (or repair times) are independent and exponentially distributed and they develop an approximating formula for the waiting time distribution function. Sherbrooke (1975) complemented their research with the derivation of an exact formula for the waiting time distribution function when the lead time is constant. Kruse (1980) and Berg and Posner (1990) develop an exact formula when the lead times are independent and have a general distribution function. Dreyfuss and Giat (2017a) characterize the functional form of this formula and develop an algorithm to find the near optimal spares allocation in a multiple-location single-echelon model. Our paper builds on Dreyfuss and Giat (2017a) and requires the reader to be familiar with its results, and therefore, we describe it in greater detail in Section 3. The inventory policy in these papers is the $(S-1,S)$ inventory policy. The waiting time distribution for different inventory review policies is developed in Kruse (1981) for an (s, S) continuous review policy, and Tempelmeier and Fisher (2010) and Kiesmüller and de Kok (2006) for an (R, s, Q) periodic review policy.

All the aforementioned papers assume that the lead times are independent. Sherbrooke (1968)'s METRIC model, however, is a multi-echelon model, for which this assumption is only an approximation, since the lead times of items returning from a higher echelon are dependent. Graves (1985) develops the VARI-METRIC model to improve the METRIC's approximation accuracy. An alternative approach is to use simulation to derive the exact values. See, for example, Shtub and Simon (1994), Caggiano et al. (2007), Caggiano, Jackson, Muckstadt, and Rappold (2009), Lee, Chew, Teng, and Chen (2008), Tsai and Zheng (2013) and Tsai and Liu (2015). Our approach is to use simulation sparingly in combination with the approximation window fill rate formulas.

The two-echelon model may be used to investigate whether pooling is beneficial in repairable-item inventory models. This problem is discussed extensively in the context of the aviation industry (e.g., Kilpi & Vepsäläinen, 2004, Wong, Cattrysse, & van Oudheusden 2005, Wong, van Oudheusden, & Cattrysse 2007, Kilpi, Töyli, & Vepsäläinen 2009). Wong et al. (2005) optimize the spares allocation problem under the assumption of full pooling, that is, spares in each location may be used by all other locations. Kilpi et al. (2009) consider different strategies of pooling; in their model, ad-hoc pooling (i.e., non-contractual agreements between two parties) is the weakest form of pooling and commercial pooling (a third party firm providing the pooling services) is the strongest form of pooling. Kranenburg and van Houtum (2009) also