European Journal of Operational Research 228 (2013) 494–503

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Production, Manufacturing and Logistics

Interval availability analysis of a two-echelon, multi-item system

Ahmad Al Hanbali *, Matthieu van der Heijden

University of Twente, School of Management and Governance, P.O. Box 217, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 24 October 2011
Accepted 1 February 2013
Available online 18 February 2013

Keywords:
Inventory
Service logistics
Markov processes
Interval availability
Reliability
Supply chain management

ABSTRACT

In this paper we analyze the interval availability of a two-echelon, multi-item spare part inventory system. We consider a scenario inspired by a situation that we encountered at Thales Netherlands, a manufacturer of naval sensors and naval command and control systems. Modeling the complete system as a Markov chain we analyze the interval availability and we compute in closed and exact form the expectation and the variance of the availability during a finite time interval $[0, T]$. We use these characteristics to approximate the survival function using a Beta distribution, together with the probability that the interval availability is equal to one. Comparison of our approximation with simulation shows excellent accuracy, especially for points of the distribution function below the mean value. The latter points are practically most relevant.

1. Introduction

Nowadays, the aftersales service business represents a considerable part of the economy and, moreover, is continuously growing (Aberdeen Group, 2005; Deloitte., 2006). Advanced capital goods such as MRI scanners, lithography systems, baggage handling systems, and radar systems, are highly downtime critical. The high criticality in these cases is due to lost production, missions that need to be aborted, patients that cannot be treated, and flights that are delayed or canceled. So the customers of these advanced goods are not just interested in acquiring these systems at an affordable price, but far more in a good balance between the resulting Total Cost of Ownership (TCO) and system productivity throughout the life cycle, including the limitation of downtime. It is often the case that the system upkeep costs during the life cycle of the system constitute a large part of the TCO. However, the core business of customers is the usage of the system and not its upkeep. Therefore, a major part of the system upkeep is preferably outsourced to the manufacturer or to an intermediate service provider that can offer a good balance between the downtime and costs. For that reason, service contracts are made between the service provider and customers. These contracts specify the services provided by the supplier with their corresponding Service Level Agreements (SLAs), such as the time between system failure and time of fault resolution, and the system availability.

The SLAs are measured over a predetermined time window, e.g., a quarter or a year. For the service providers, it is essential that the service levels are attained, because in some cases penalties apply if an SLA target is violated. In case of a large scale service contract (the average performance over many systems is measured), the average performance should meet the target. If the number of systems covered by a contract is relatively small, we have inherent statistical variability and we need an additional buffer in performance to assure that the probability of not meeting the SLAs over the time window is still acceptable. We encountered such a situation at Thales Netherlands, a manufacturer of naval sensors and naval command and control systems. There, a service contract typically covers a few systems only. In the literature, this issue is usually neglected. In this paper, we are mainly interested in the logistical delay due to the unavailability of spare parts, since this is the basis of current service contracts at Thales Netherlands. Moreover, the focus will be on SLAs that are based on the system availability during a predetermined period of time.

In service parts logistics there is usually a tradeoff between the cost involved in keeping the stocks very close to the customers sites or at a central depot, which can supports multiple customers at the same time. Due to the risk pooling effect, it is more desirable for a service provider to position the stocks of spare parts centrally. However, having a strict SLA, e.g., 99% availability in a quarter, forces the service provider to move some spare parts closer to the customer sites. In addition, in order to reduce the system downtime and its critical consequences, the repair of a failed system is usually done by replacing the failed part with a new part. The failed part is sent to the repair shop, i.e., the inventory is managed using one-for-one replenishment, so an $(s-1, s)$-policy. This policy is justified by the fact that most parts are slow movers for which a replenishment order of size one is usually (near) optimal.

* Corresponding author. Tel.: +31 534894033.
E-mail address: a.alhanbali@utwente.nl (A. Al Hanbali).

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2013.02.009

Sherbrooke (1968) was among the first to tackle the spare part optimization problem. He proposed the METRIC model that is based on the maximization of system availability subject to a constraint on the invested budget in spare parts. The main decision in METRIC is how much to keep in stock at each of the locations in the supply network. The METRIC model provides good approximations for multi-echelon spare part networks, especially in case of a high availability. Graves (1985) and Slay (1984) extended the METRIC model and proposed an improved approach called the VARI-METRIC. We note that the VARI-METRIC model is the approach used in most commercial software tools on spare parts optimization.

It is worth to mention that both METRIC and VARI-METRIC and most spare parts management theory are based on finding an optimal balance between the initial spare part investment and the steady state system availability, i.e., the fraction of time the system is operational during a very long (infinite) period of time. However, in practice we often see that the agreed upon availability SLA is the average availability during a finite period, e.g., month, quarter, or year. Moreover, if the availability during a period of time is lower than a specific percentage, penalty rules apply. This motivates us to analyze the availability during a finite period of time, the so-called interval availability that is defined in reliability theory as follows, see, e.g., (Nakagawa & Goel, 1973):

Definition. The system interval availability is defined as the fraction of time a system is operational during a period of time $[0,T]$.

Note that in (Barlow, Proschan, & Hunter, 1965; Hosford, 1960) the interval availability is defined as the *expected* fraction of time a system is operational during $[0,T]$. To avoid confusion in this paper and according to the previous definition the interval availability is a random variable that has a distribution. In addition, this probability distribution has a finite support between zero and one with probability masses at the points zero and one: There are strictly positive probabilities that (i) an operational system will not face any lack of spare parts during $[0,T]$, and (ii) a failed system waiting for a specific spare part will not be repaired by replacement during $[0,T]$. In practical instances, the first probability will be significant, and the second probability will be close to zero.

Our main contribution in this paper consists of the following points:

- We propose a computational efficient and accurate approximation for the interval availability of a multi-item system supported by a two echelon supply network. More specifically, our approximation is accurate in the practical case of systems with high average availability.
- As part of this approximation, we derive in closed-form the variance and the third moment of the cumulative sojourn time in a subset of states of Markov chain in a finite interval. In principle, we can also derive all the higher moments using the same approach.
- Using simulation we show that the survival function of the interval availability is not very sensitive to the order-and-ship time distribution at the points of survival function that are below the expected availability. This justifies our Markovian approach, specifically, the assumption of exponential order-and-ship times.

The paper is organized as follows. In Section 2 we briefly review the related literature. Section 3 describes our model and the assumptions used to analyze the interval availability distribution of a two echelon, multi item supply network. In Section 4 we report our approximation where our key results are reported in a set of Theorems. In Section 5 we validate our approximation using simulation and evaluate the impact of the order-and-ship time on the interval availability. Finally, in Section 6, we conclude the paper and give some directions for further research.

2. Related literature

In this section we shall review the existing literature on interval availability. Takács (1957) is among the first to analyze the interval availability distribution function of an on–off stochastic process. Takács result is in the form of an infinite sum of terms, each consisting of multiple convolutions. This result is hard to compute numerically. van der Heijden (1988) approximates the interval availability distribution using two-moment approximations for the on and off periods, which yields accurate results within small computation times. Another approximation based on fitting the *approximated* first two moments, the hundred percent, and the nil probability of the interval availability in a Beta distribution is proposed in (Smith, 1997). For an on–off two states Markov chain the first two moments of the interval availability are derived exactly in (Kirmani & Hood, 2008). We note that in all these previously mentioned studies the underlying assumption is that the on periods are independent and the off periods are independent, moreover, all the on and off period are independent of each other, i.e., the on–off process can be represented by a renewal process.

De Souza e Silva and Gail (1986) derive in closed-form the cumulative sojourn time distribution in a subset of states of a Markov chain during a finite period of time. The subset of states can, for example, represent the operational states of a system. Therefore, the division of the cumulative sojourn time by the period length gives right away the system interval availability. We note that computing the full curve of the interval availability distribution using the method of De Souza e Silva and Gail (1986) or its improved version in (Rubino & Sericola, 1995) is time consuming. Carrasco (2004) proposes an efficient algorithm to compute the interval availability distribution for the special case of the systems which can be modeled by an *absorbing* Markov chain. In the latter three papers the renewal assumption of the on–off process is not necessary.

In this paper, we propose a numerically efficient approach to compute the distribution function of the interval availability. Our approach builds on the result of De Souza e Silva and Gail (1986) extensively in order to compute in closed-form the first two moments of the interval availability. These two moments have not been derived previously in the literature for a Markov chain with more than two states. Moreover, we follow a similar approach to (Smith, 1997) to approximate the interval availability by a Beta distribution using the first two moments in addition to the hundred percent probability of the interval availability.

Finally, we note that the analysis of a service level over a finite period of time is not only of interest in reliability theory, but also in inventory management of fast moving products where demand is typically modeled by a Normal distribution. See, e.g., (Banerjee & Paul, 2005; Chen, Lin, & Thomas, 2003) in which the interest is on the expected fill rate over a finite period of time T for a single site, single item system. In these papers, it is proven that the expected fill rate over a finite period is larger than over the infinite period case. Thomas (2005) evaluate the impact of T and the demand distribution on the fill rate distribution over T. In the latter paper, simulation is used due to the difficulty in explicitly computing the fill rate distribution during T. Tactical decisions on stock level to meet the time-based SLA in the case of multi-echelon, single item scenario are considered in (Cohen, Kleindorfer, & Lee, 1986) and for the multi-item scenario in (Ettl, Feigin, Lin, & Yao, 2000). The restriction in the analysis is that the time period should be equal to the supply lead time of the part. More recently, the model in the latter two papers is extended and a scalability analysis is added in (Caggiano, Jackson, Muckstadt, & Rappold, 2007).