2008

Multi-Echelon Repairable Item Inventory System with Limited Repair Capacity under Nonstationary Demands

Hoong Chuin LAU
Singapore Management University, hclau@smu.edu.sg

Huawei SONG

DOI: https://doi.org/10.1504/IJIR.2008.019209

Multi-echelon repairable item inventory system with limited repair capacity under nonstationary demands

Hoong Chuin Lau

School of Information Systems,
Singapore Management University,
80 Stamford Road, 178902, Singapore
E-mail: hclau@smu.edu.sg

Huawei Song*

School of Engineering,
Rensselaer Polytechnic Institute,
110 Eighth Street Troy, NY 12180, USA
E-mail: songh2@rpi.edu
*Corresponding author

Abstract: Classical multi-echelon repairable item inventory models are based either on steady-state analysis or infinite repair capacity, which may not work well in situations when the demand is nonstationary, or repair capacity is limited. In this paper, we propose an analytical model for evaluating system performance that works well under limited repair capacity and nonstationary demands. Following the METRIC methodology, we then develop an optimisation algorithm to solve the corrective maintenance problem in military logistics. Experimental results show that our approach yields good solutions efficiently. This work has also resulted in a software that has been field-tested by a military organisation.

Keywords: multi-echelon; repairable item; inventory; queueing; maintenance; military logistics.

Reference to this paper should be made as follows: Lau, H.C. and Song, H. (2008) 'Multi-echelon repairable item inventory system with limited repair capacity under nonstationary demands', *Int. J. Inventory Research*, Vol. 1, No. 1, pp.67–92.

Biographical notes: Hoong Chuin Lau is an Associate Professor at the Singapore Management University School of Information Systems. He holds a concurrent appointment as Director of Defense Logistics at The Logistics Institute Asia Pacific. His interests are computational methods and models for logistics and supply chain management. He has worked extensively on defence logistics problems, and he has developed a number of tools and systems, which have been field-tested and deployed.

Huawei Song is a PhD student in Decision Science and Engineering Systems in Rensselaer Polytechnic Institute in the USA. His current research interest is in financial engineering, risk management and optimisation, particularly in finance, economics and logistics. He obtained his Master's Degree in Computer Science from the National University of Singapore in 2002 and worked as a Research Engineer in The Logistics Institute – Asia Pacific (TLI-AP) for

four years. When he was in TLI-AP, he was working on Jaguar project, optimisation of resources in multi-echelon inventory systems.

1 Introduction and motivation

Advanced systems, especially military systems such as aircrafts, have expensive complex structures that break down because components are either worn out or damaged during operations. To support high operational readiness (or availability), sufficient quantities of spare components (called Line Replaceable Unit or LRU) and maintenance resources (comprising repair manpower and tools) are required to sustain demands arising from LRU breakdowns (or failures). However, since spares and resources are costly, consume space, and become obsolete over time, there is a trade-off between cost and availability. The goal of the planner is to sustain the life cycle of systems with respect to cost and availability.

The importance of spare parts management has increased in the past decades owing to the increasing value of service part inventory investment. As an example, a survey on computer manufacturers conducted by Cohen et al. (1997) reveals that service parts inventory investment take up a significant 8.75% of the value of product sales.

The motivation of this work arises from the task of designing an automated tool for optimising the sustenance of military systems[1] for near-future mission planning purposes. To sustain a military system, the planner should decide how many units of spares should be bought over time, and what repair capacity should be set aside for the repair of repairable parts throughout its lifetime. Unfortunately in real life, it is hard for planners to make such decisions over the long term owing to the following reasons. First, the demand (i.e., breakdown) rate varies as time because of system utilisation rates that fluctuate from season to season. This situation is even more acute in the military context, since military systems usage varies from one mission to another, and transits rapidly from peace to wartime. Most academic and even commercial offerings assume that the demand is given by a *stationary* Poisson process with a mean such as the annual demand rate. In reality, however, point estimates based on mean values will cause high errors at certain time points. Another reason peculiar to the military context – and increasingly so in today's commercial context as well – is that the (mission) planning period is progressively getting shorter, to the extent that the underlying inventory system may not converge even to steady state. As an example, the lifetime of an aircraft is typically 10–30 years, whereas it is necessary to make decisions and predict the system behaviours based on utilisation rates, which are available on, say, 6-monthly basis. Furthermore, utilisation information is usually probably not available for a much longer period. Finally, military systems usually exhibit low failure rates, which make it difficult even to forecast the average rate in the first place.

Hence, the planner is interested to predict and plan for the near future (say 6-monthly), during which exercise or mission schedules (and hence utilisation rates) are known. Furthermore, on a rolling operating horizon basis, the planner will continue predicting and planning based on the existing system performances at hand when information becomes available. During wartime, this allows the planner to plan the next operation/mission in response to dynamically changing demand information such as

combat damage. There is a need to predict how the inventory performance such as availability fluctuates over time given certain spares and repair resources allocation. This will help planners consider whether to buy extra spare parts or hire extra manpower in the next period.

Motivated by the above requirement, this paper is concerned with a 2-echelon repairable item inventory system under nonstationary Poisson demands (Nelson, 1995)[2] *and* limited repair capacity. We present how to evaluate the system performance of a given spares and resource allocation configuration. Using that evaluation model, we then devise an efficient algorithm that generates optimal solutions for the quantities and joint allocation of spares *as well as repair resources* to sustain time-varying demands. We call this the *corrective maintenance problem*.

At this early stage, it is important to understand what we mean by time-varying demands. Our view of time-varying demands is based on our project experience with a military organisation. The demand rate of a given LRU item is derived by and dependent on several factors. Some of the common ones are the system utilisation rate (i.e., the percentage/fraction of system usage varying over a certain time interval), the Mean Time Between Failures (*MTBF*) of the item (i.e., the expected value of time duration between two consecutive failures), the number of operating military systems, and the number of LRU items of the given type per military system. Of particular interest is the system utilisation rate. As an illustration, we borrow a real-life military system and give an example of her 6-month utilisation rate as follows: 0.0486 (0–1440 h), 0.0833 (1440–2160 h), 0.5833 (2160–2520 h), 0.3333 (2520–3240 h), 0.5 (3240–3600 h) and 0.25 (3600–4320 h). Observe that during the entire operating period of six months, the utilisation rate varies frequently and drastically from phase to phase. It is clear that a stationary Poisson process cannot accurately model such a demand pattern. Instead in this paper, we approximate this demand pattern by a *nonstationary* Poisson process with varying mean demand rate over time (i.e., the demand rate is now a function of time t, denoted $\lambda(t)$). We will use the term 'nonstationary' interchangeably with the term 'time-varying' or 'time-dependent'.

This paper is organised as follows. A literature survey is provided in Section 2. In Section 3, we describe the problem formally. Section 4 presents an analytical model to evaluate the system performance given current spares and repair resources allocation. In Section 5, we present a nonstationary multi-class finite-server queuing system, which is a core computational problem underlying our model and then present how to compute the expected number of each class of customers in the queuing system at any time efficiently. An optimisation algorithm based on the ideas proposed in Sections 4 and 5 is then developed in Section 6 to solve the corrective maintenance problem. Experimental results are shown in Section 7 followed by conclusion and future work in Section 8.

2 Literature review

Optimisation of multi-echelon repairable inventory systems has been an active area of research over the last 30 years. Much of the research has focused on steady-state spare allocation and very little attention has been placed on nonstationary demands or the issue of finite resources. In this work, we take interest in four key parameters: the underlying echelon network, indenture structure, repair resource capacity and demand distribution. Accordingly, we provide a classification of some major analytical multi-echelon models

based on the key parameters above. Following classical scheduling terminology, we classify these models in the form of a/b/c/d, annotated as follows. The first field is the echelon structure where **1** denotes single-echelon model, **2** denotes 2-echelon model and **m** denote multi-echelon model. The second field is indenture structure where **1** denotes single-indenture (items have only LRUs) while **2** denotes two-indenture (items have both LRUs and their sub-components) and **m** denotes multi-indenture. The third field is capacity of repair resources, which is **∞** for infinite repair capacity and **S** for limited repair capacity with S servers. The default value is **∞** if this field is blank. Finally, the fourth field is demand process, which is λ for stationary Poisson distribution or λ_t for time-dependent Poisson distribution. The default value is λ if this field is blank. Table 1 summarises key classical works according to our classification.

Table 1 Classifications of models

Problem	*Reference*
2/1	Sherbrooke (1968) (METRIC)
2/1	Graves (1985)
2/m	Sherbrooke (1986) (VARI-METRIC)
m/m	OPUS9 (1992) and OPUS10 (1998)
2/1/S	Díaz and Fu (1997)
2/1/S	Alfredsson (1997, 1999) (OPRAL)
3/3/S	Sleptchenko et al. (2002, 2003)
2/1/S/λ_t	Jung (1993)
2/1/∞/λ_t	Slay et al. (1996)
m/3/∞/λ_t	Isaacson and Boren (1988) Dyna-METRIC

The following subsections trace the evolution of the analytical models, followed by a brief discussion of simulation and queuing models.

2.1 Analytical models

2.1.1 METRIC

Multi-Echelon Technique for Recoverable Item Control (METRIC) by Sherbrooke (1968) is the pioneer study for the majority of multi-echelon repairable-item models that follow. METRIC assumes that there is infinite repair capacity (implying no queue at the depot), hence allows repair times to be independent of the number of items in repair. The failures at bases are assumed to be Poisson and hence the number of items in the pipeline of depot also follows a Poisson distribution. Under the assumption of first-come-first-served replenishments from the depot to the bases, the distribution of the number of items in the base pipelines is also approximated to be Poisson. Under this environment, METRIC solves the spares allocation problem elegantly and efficiently by marginal analysis.

2.1.2 METRIC extensions

Arguably, METRIC is a simplistic model and during its implementation, it was found that the Expected Number of Backorders (EBO) computed was often underestimated due to the use of Poisson distributions. Graves (1985) proposes to model the distribution of the number of items in the base pipelines by a negative binomial distribution, i.e., it uses the variance parameter to reduce the gap. The improvement comes from the observation that the variance-to-mean ratio must be one under Poisson distribution, whereas it is usually greater than one in practice. Under Graves' model, both mean and variance of backorders are calculated and the probability distribution is chosen based on the variance-to-mean ratio. It has also been proved that Graves' model performs equivalently to METRIC when the depot stock level is zero. When the depot stock is not equal to zero, empirical results show that Grave's model produces more than 99% accuracy in spares allocation whereas METRIC achieves around 89% accuracy. Sherbrooke (1986, 1992) proposes VARI-METRIC that captures variance based on Graves' model. This model is interesting in that if the variance-to-mean ratio is of the pipeline greater than one, negative binomial distribution will be adopted. If it is equal to one, Poisson distribution will be adopted. If it is less than one, binomial distribution will be adopted. The OPUS9 (1992) and OPUS10 (1998) are METRIC-based spares optimisation software tools developed commercially by Systecon. Besides adopting the structure and assumptions of METRIC, the tool provides additional features, e.g., the user has the flexibility to specify problem scenarios and order policies.

2.1.3 Limited repair capacity

The models discussed hitherto assume infinite repair capacity, which is often an unrealistic assumption in industrial contexts. More specifically, such an assumption will underestimate the quantity of spare parts needed in systems with high repair facility utilisation. Díaz and Fu (1997) first relax this assumption by considering limited repair facilities at the depot. They consider the setting where all failed LRUs are repaired at the depot and propose results and approximations based on queuing theory for three cases – where the queue at the depot follows *M/M/s* single-class model, *M/G/s* single-class model and *M/G/s* multi-class model. For the *M/M/s* single-class model, the failure follows a Poisson process and repair time follows an exponential distribution with limited repair facilities (servers). Based on single-class, different types of failed LRUs will require different types of servers and only one unit of the required type of server. The mean and variance of the number of items in the repair facility, both in queue and in repair, are calculated using standard *M/M/s* queuing theory. The model is then extended to *M/G/s* where the repair process follows a general distribution. This is further extended to a multi-class model that allows each type of server to be used to repair multiple types of LRUs. Díaz and Fu (1997) provide an aggregation–disaggregation approach to calculate the first two moments of per-class number in queue and repair. Unfortunately, the variance of per-class number in queue and repair pipeline is derived only for the single-server multi-class queue model due to analytical complexity.

This line of work has been extended recently in several interesting ways. Sleptchenko et al. (2002, 2003) use a more general multi-class multi-server queuing model for the repair shop under steady state when the repair capacity is given. Perlman et al. (2001) use congestion externalities to set expediting repair policy to choose

either the repair mode with a normal repair time or the one with an expedited repair time. However, to use exact methods, they restrict themselves to a single repair capacity shop. Kim et al. (2000) extend previous results to the system where spares allocated at the bases as well as the depot.

2.1.4 Joint spares and resource allocation

The models discussed so far only consider the spare allocation problem addressing the question "how many spares to stock and where to put them". The number of repair facilities is assumed to be fixed and the optimal allocation of repair resources has not been considered. Alfredsson (1997, 1999) proposes OPRAL (an offshoot of OPUS) that tackles the joint spares and resource allocation problem within a single model. It considers the question of "how much repair capacity is needed and where to allocate resources" in addition to the spares allocation problem. It assumes that each failed LRU requires only one resource type but different LRUs may share a common repair resource type. This assumption implies that LRUs can be partitioned into disjoint *resource groups*, each of which contains the LRUs that require that particular resource type. The queue within a resource group is modelled as *M/M/s* to calculate the expected waiting time for an available resource. Poisson distribution is used primarily rather than negative binomial distribution, but some passing remarks were made on the use of negative binomial distribution to improve the fidelity of the model.

More recently, Zijm and Avsar (2003) consider optimal resource allocation under capacitated two-indenture models, but within only a single site. Slepchenko et al. (2003) present a procedure for joint optimisation of spares and repair capacities, especially for noninteger repair capacities. Both these models consider finite repair capacity under steady state (i.e., Poisson demands).

2.1.5 Time-varying demands

All the above models are steady-state models, which work well when demand follows a stationary Poisson distribution, i.e., the demand rate is constant over time. Unfortunately, many repairable items have long lifecycles and hence the demand rates will inevitably change with time. In a time-varying demand situation, these models will not produce accurate results. Jung (1993) first presents a methodology for a recoverable inventory system with time-varying demand using discrete event simulation. The echelon structure is based on METRIC with limited repair capacity at the depot, except that the demand rate tends to decrease in successive periods. Thus, the repair process at the depot is modelled as a nonstationary *M/M/s* system. The expected number of items in queue and repair at the depot is time-dependent due to nonstationary Poisson process. Jung (1993) implements the SIMAN system for computing this time-dependent value with the empty queue condition at the beginning. Given a fill rate target, it presents a method to determine the stock level at a certain given time point. Unfortunately, the limitation is that only a single item type is allowed, and the method does *not* perform optimisation.

Slay et al. (1996) propose an aircraft sustainability optimisation model that can handle problems with time-dependent demand rates but under infinite resources. The underlying echelon structure is that a depot only supports a base, and the failure at the base is a nonstationary Poisson process whose mean value varies with time.

As an optimisation model, it only considers spares allocation but not resource allocation, and it investigates the objective and spare allocation only at *specific* time points of interest. In this model, the failure rate need not decrease with time (as required by Jung, 1993) – it is high during wartime and low in peacetime. The repair time and shipment time may or may not be time-dependent. The expected number in the pipeline will be calculated first by using integration and then the EBO will be calculated at the certain time point of interest.

RAND Corporation has developed a proprietary system called Dyna-METRIC to serve the US Department of Defense. The Dyna-METRIC series are capability assessment models designed to explore ways to improve wartime logistics support to aircraft. They can solve many problems including nonstationary demands and cannibalisation to assess the effects of wartime dynamics and projects operational performance measures. Version 5 (Isaacson and Boren, 1988) is the latest analytical model to-date in which logistics support system is assumed to be 5-echelon and 3-indenture. Version 5 has its limitations. First, it assumes that the aircrafts deployed at each base are identical, i.e., it does not deal with different item types. Second and unfortunately, like Slay et al. (1996), it provides a steady-state solution to a time-dependent problem.

2.2 Simulation models

In this subsection, we briefly discuss three influential simulation models, namely Dyna-METRIC (Version 6), SPAR and Pyke (1990).

While Version 5 of Dyna-METRIC is an analytic model based on dynamic form of Palm's theorem, Version 6 is a Monte-Carlo simulation tool as an answer to limitations imposed by analytic models, such as infinite capacity. It is a 3-echelon, 2-indenture model that accommodates interesting features such as lateral supply between bases, lateral repair, information lags and exception reporting. It allows items to have priority to be repaired not only based on FCFS scheduling policy. Although superior to analytical versions in repair process, version 6 has its own limitations. For instance, it does not compute spares requirements because the equations on spares allocations are unavailable in the simulation. It has to draw support from analytical model. Another limitation is that simulation model is usually very slow when compared with executing the counterpart analytical model. SPAR (http://www.clockwork-group.com) is a commercial Monte-Carlo simulation tool that deals with the problem of time-dependent demands. At the point of writing, SPAR requires external FORTRAN programming to perform the role of repair resource allocation. We also observe that optimisation is slow because it entails running simulation many times. Pyke (1990) presents a simulation study for repairable electronic equipment used by military aircrafts. This model covers a 3-echelon system: a repair depot, a stockpile of repair parts and a set of bases. It considers priority rules for allocating repaired items to bases and sequencing items at the repair depot. It also considers the importance of the initial allocation of a fixed amount of stock, and the lateral transhipment that occurs only when it is possible to fill all the backorders of a specific base. The optimisation is performed taking into account three decisions: repair rule, distribution rule and where the initial stock of spares is allocated.

2.3 Queuing models

A line of work in queuing theory is concerned with approximations under time-dependent arrival and service rates. One of the well-known approximations is the Pointwise Stationary Approximation (PSA), which assumes pointwise stationarity in time and approximates long-run average performance measures (see Green and Kolesar, 1991, 1997; Green et al., 1991). Another approximation is the 'closure approximation', which employs negative binomial distribution to approximate the time-dependent measures (Rothkopf and Oren, 1979). In this paper, as part of the effort in evaluating system performance, we will be studying and adapting the results of Rothkopf and Oren (1979).

3 Problem definition

In Section 3.1, we provide the scope of our problem. Section 3.2 gives the notations used; and Section 3.3 gives a formal problem formulation.

3.1 Problem scope

In our environment, there is a single depot that supports a number of *bases* where military systems are deployed. Each military system is composed of multiple items (LRUs), which are assumed to be connected in series. For simplicity, we assume that all systems are identical. Spares can be allocated at both bases and the depot, whereas repair resources are only allowed to be allocated at the depot. For simplicity, repair resources are categorised into resource types. When an LRU at a base fails, the system will be grounded at the base and the failed LRU will be removed and replaced by a spare unit if it is available at the base. Otherwise, there is a backorder at the base and the entire system has to wait until a spare is available. The failed and removed LRU will be delivered to the depot where finite repair resources are allocated. If the required repair resources are available, this LRU will be repaired. Otherwise, it has to wait for repair. When repair is completed, the 'good' LRU will be placed in the depot stock to meet future demands.

In this paper, the following assumptions are made:

1 All LRUs must and can be repaired at the depot.

2 Continuous resupply, i.e., an LRU can be shipped between the depot and bases without delay at any time (i.e., there is an infinite number of transporters). However, transportation time is incurred, and is known as order-and-ship time.

3 $(s - 1, s)$ inventory policy is applied for all LRUs at all sites because LRUs are usually expensive with low failure rate.

4 The repair time of an item follows an exponential distribution.

5 Each LRU requires exactly one repair resource type. Different types of LRUs may compete for the same repair resource type, and resources are assigned to contending LRUs according to the FCFS policy.

6 FCFS replenishment from the depot to bases.

7 There is no lateral supply, i.e., no supply or shipment between bases.

The justification for these assumptions is given as follows. First, contrary to Sleptchenko et al. (2002, 2003) where each location consists of a repair shop, we assume that all items must and can be repaired at the depot. Our model can be extended easily to the one in Sleptchenko et al. (2002) so long as the repair probabilities are mutually exclusive, as assumed in Sleptchenko et al. (2002). Assumptions (2)–(4) are standard assumptions in the existing literature. Assumption (5) is also well assumed where one resource can be one team. As for Assumption (6), although Pyke (1990) shows by simulation that priority queuing has a positive impact on the system performance, we adopt the FCFS discipline in replenishment as done in Sleptchenko et al. (2002). Assumption (7) is justified as stated in Sleptchenko et al. (2002) since lateral supply has only significant impact for low fill rates at downstream locations. These assumptions imply that the various pipelines can be estimated fairly accurately by time-varying Poisson distributions, and hence our analysis in this paper will be based on mean values only.

We are concerned with two issues. First, we evaluate the time-dependent system performance of a fixed spares and resource configuration for the $2/1/S/\lambda_t$ problem. To be consistent with the METRIC and other literature, we adopt the *Expected Number of Backorders* (*EBO*) as the measure of system performance. Arguably, one may also use a related metric such as *Ao* (availability). Typically, *EBO* and *Ao* are inversely proportional to one another, and one such relationship is given in Lau et al. (2006). Second, we are also interested in solving the *joint* spares and resource optimisation problem for $2/1/S/\lambda_t$. Without loss of generality, our approach can readily be extended easily to more than two echelons and more than one indenture. Compared with the existing literature, this can be viewed as an extension of Alfredsson (1997, 1999) in tackling time-dependent demands, or extension of Isaacson and Boren (1988), Jung (1993) and Slay et al. (1996) in tackling finite resources from a non-steady-state optimisation perspective.

3.2 Notations

We adopt and extend the notations of those in Alfredsson (1997, 1999) and Sherbrooke (1992). We use j to index the sites, where $j = 0$ for the depot and $j = 1, \ldots, J$ for the bases. LRU types are indexed by $k = 1, \ldots, K$, while repair resource types are indexed by $g = 1, \ldots, G$. Time is indexed by $t = 1, \ldots, T$ where T is the operating horizon. The following notations are used:

$MTBF_k$: Mean Time Between Failures of LRU k

TAT_k: Mean repair time of LRU k (often called turnaround time in practice and some literature)

OST_k: Constant order-and-ship time for LRU k from the depot to the base

$Nsys_j$: Number of military systems deployed at base j

QPM_k: Quantity of LRU k contained in a military system (often called quantity per mother-item)

$UR(t)$: System utilisation rate at time t (system utilisation rates are assumed to be the same across bases)

Cs_k: Unit cost of spare LRU k

Cr_g: Unit cost of repair resource g

s_{0k}: Number of spare units of LRU k at the depot

s_{jk}: Number of spares of LRU k at base j

r_g: Number of repair resource g at the depot.

Since the demand is time-dependent, EBO will inevitably be time-dependent and is denoted by $EBO(t)$, which we define as the sum of expected backorders over all LRUs and all bases *at time t*.

3.3 Problem formulation

The first problem of evaluating system performance is defined as: given a fixed spares and resource allocation configuration (s_{0k}, s_{jk}, r_g), compute $EBO(t)$ analytically. Our proposed model will be discussed in Sections 4 and 5.

The second problem is an optimisation problem driven by cost and system performance. The cost model we consider is a function of the investment cost incurred by spares and resources. The investment cost for spares is straightforward, and is computed by the unit costs multiplied by the number of spare units purchased. The investment cost for resources, however, is a little tricky. As spares are purchased and circulate in the system for a number of years, we should calculate a kind of 'purchasing price' for repair men[3] using the net present value of all the cost for a repair man throughout the system lifecycle. These expenditures include wages, taxes and social premiums, housing, education, tools, etc., Sleptchenko et al. (2003). We will use the terminology LSC (standing for *Life Support Cost* in OPUS9 (1992) and OPUS10 (1998)) as the notation for total cost, where

$$LSC = \sum_{k=1}^{K} C_{s_k} \left(\sum_{j=0}^{J} s_{jk} \right) + \sum_{g=1}^{G} C_{r_g} \times r_g . \tag{1}$$

Let $maxEBO = \max_{t\in[0,T]} EBO(t)$. Given a budget amount B, one problem is to find an allocation of spares and repair resources (i.e., deciding the values of (s_{0k}, s_{jk}, r_g)) that minimises *maxEBO* while not exceeding the budget (i.e., min *maxEBO* s.t. $LSC \le$ B). Conversely, we wish to find a minimum-cost spare and repair resource allocation such that the EBO *at any time* within the operating horizon will not exceed a specified target $E_{\max}$, (i.e., min *LSC* s.t. $maxEBO \le E_{\max}$).

Note that both these optimisation problems are generalisations of the knapsack problem, which renders them NP-hard.[4] The good news, however, is that planners are usually not interested in the optimal allocation point with respect to a specific budget or a target EBO, but rather the problem is to combine the two problems as one by seeking a Cost/Effectiveness (or C/E) curve where each point on the curve is an optimal allocation associated with a cost and EBO value (see Figure 4). This is known as the *corrective maintenance problem* in this paper. It is well known that marginal analysis, together with convexification, provides an efficient polynomial-time solution approach to solve this problem (Sherbrooke, 1992). The idea of applying marginal analysis to plot the C/E curve will be adopted in our optimisation algorithm presented in Section 6.

4 Evaluating system performance

In this section, we present our analytical model for evaluating the system performance $EBO(t)$. The following is a list of intermediate (i.e., state) variables used:

$EBO_{0k}(t)$: EBO of LRU k at the depot at time t

$EBO_{jk}(t)$: EBO of LRU k at base j at time t

$\lambda_{0k}(t)$: Demand rate of LRU k at the depot at time t

$\lambda_{jk}(t)$: Demand rate of LRU k at base j at time t

$RP_{0k}(t)$: Random variable representing number of LRU k in the **depot** repair pipeline at time t

$BP_{jk}(t)$: random variable representing number of LRU k in the pipeline of **base** j at time t

$OSP_{jk}(t)$: Random variable representing number of LRU k in the **order-and-ship** pipeline to base j at time t

$f_{jk}(t)$: Fraction of LRU k at base j contributing to the EBO at the depot at time t.

As in Sherbrooke (1992), the notation $EBO(s|\lambda)$ is used to denote value of EBO given stock level s when the mean pipeline is λ. Following standard probability, this quantity is computed as $\sum_{x>s}(x-s)\Pr\{X=x\}$ where X is the pipeline random variable with mean λ.

In the following, we will present a bottom-up derivation of the EBO function.

First, we compute the demand rate. Similar to the assumption in METRIC and Perlman et al. (2001),[5] the demand rate of LRU k at base j is computed by definition as follows:

$$\lambda_{jk}(t) = \frac{UR(t)}{MTBF_k / QPM_k} \times Nsys_j. \tag{2}$$

Hence, by aggregating demands, we can compute the depot demand rate as:

$$\lambda_{0k}(t) = \sum_{j=1}^{J} \lambda_{jk}(t). \tag{3}$$

Next, we compute $RP_{0k}(t)$, the depot pipeline of LRU k. This is a Poisson random variable with mean equal to the expected number of LRU k in the repair facility, consisting of those in queue (waiting for repair resources) and in process (being repaired), at time t. Let $N_k(t)$ be a random variable representing this quantity. Hence,

$$E[RP_{0k}(t)] = E[N_k(t)]. \tag{4}$$

Unfortunately, the quantity $N_k(t)$ cannot be easily computed under finite capacity, and *Section 5* will provide a method to estimate this quantity using a queuing model.

Given the stock of LRU k at the depot s_{0k}, the depot EBO at time t is, by definition:

$$EBO_{0k}(t) = EBO(s_{0k} \mid E[RP_{0k}(t)]). \tag{5}$$

Having computed $EBO_{0k}(t)$, we can now derive the base pipeline $BP_{jk}(t)$. By definition, this is a sum of the order-and-ship pipeline (i.e., those in transportation) by time t for

those items having been shipped out from depot by time *t*-OST, plus the portion of the depot repair pipeline attributed to the base, which cannot be shipped out by *t*-OST. From the splitting theorem for nonhomogeneous Poisson process (Slay et al., 1996), we know that a sum of Poisson processes is still a Poisson process. Hence, $BP_{jk}(t)$ is a Poisson process, and by the linearity of expectation, its expected value is given as:

$$E[BP_{jk}(t)] = E[OSP_{jk}(t)] + f_{jk}(t - OST_k)EBO_{0k}(t - OST_k). \tag{6}$$

Having computed $BP_{jk}(t)$, we can now compute the EBO at base, $EBO_{jk}(t)$. Given the stock of LRU *k* at base *j* s_{jk}, the EBO of LRU *k* at base *j* at time *t* is given by:

$$EBO_{jk}(t) = EBO(s_{jk} \mid E[BP_{jk}(t)]). \tag{7}$$

And finally, our system performance measure $EBO(t)$ is:

$$EBO(t) = \sum_{k=1}^{K} \sum_{j=1}^{J} EBO_{jk}(t). \tag{8}$$

4.1 Derivation of intermediate variables

It remains to show how the intermediate variables $OSP_{jk}(t)$ and $f_{jk}(t)$ are derived. Under the assumption that the number of components in the pipeline follows a *time-dependent* Poisson distribution, the expected number of demands in the pipeline at time *t* can be computed by a dynamic form of Palm's theorem, presented in Carrillo (1991) by relaxing the arrival process and service time distribution assumptions:

Theorem 1 (Carrillo, 1991)**:** *Suppose we have nonhomogeneous Poisson arrival with intensity function* $\lambda(t) \geq 0$ *for* $t \geq 0$, $\lambda(t) = 0$ *otherwise, and nonstationary service distribution G. Then, the number of arrivals undergoing service at time t has a Poisson distribution with mean*

$$\Lambda(t) = \int_0^t (1 - G(s,t))\lambda(s)\mathrm{d}s \tag{9}$$

where the random service time Y at time t has the distribution $P\{Y(t) \leq y\} = G(t, t + y)$.

Hence, under the assumption of constant transport time, Theorem 1 shows that $OSP_{jk}(t)$ is indeed a *Poisson* random variable with mean equal to

$$E[OSP_{jk}(t)] = \int_{(t - OST_k)^+}^{t} \lambda_{jk}(s)\mathrm{d}s \tag{10}$$

because by time *t*, all items shipped out within time slot (*t-OST*, *t*) are still in the order-and-ship pipeline since it takes time OST to transport items while those shipped out before *t-OST* have been out of the transport pipeline.

Next, we will show how to compute $f_{jk}(t)$, i.e., how to distribute EBO at the depot among the bases. The trick is to distribute it according to the proportion of their respective demand rates. This is a good approximation due to two reasons. First, we assume FCFS replenishment policy from depot to bases. Second, we assume that system utilisation rates are the same across all bases implying that demand rates among bases vary synchronously over time. These two assumptions imply that the waiting times for an available depot spare are the same across all bases. Hence,

$$f_{jk}(t) = \frac{\lambda_{jk}(t)}{\lambda_{0k}(t)}. \tag{11}$$

With these quantities defined, EBO can now be evaluated according to equations (6)–(8).

5 Nonstationary multi-class finite-server queue

The previous section shows that the EBO computation involves the calculation of the expected number of failed LRUs in the repair facility (i.e., $N_k(t)$ for all LRU types 1, …, k), which will be presented in this section. Note that while Palm's theorem is readily applicable in case of infinite capacity, under finite capacity, failed LRUs may need to wait for resource availability and hence exhibit a queuing system behaviour. Furthermore, when compared with classical queuing models such as $M/M/s$, we have multiple classes of customers (LRUs), each of which has its own arrival and service rates, where the arrival process is a nonstationary Poisson process.

To be consistent with queuing theory terminology, a repair resource is equivalently called a server and an LRU is called a customer. Under our assumption that each LRU type requires exactly one repair resource type, LRUs (or more precisely LRU types) can be partitioned into disjoint resource groups, each of which consists of all LRUs competing for the certain type of resource. The scheduling policy within each resource group is FCFS, i.e., all demands competing for the same resource will wait in a single queue until one of the identical parallel repair resources is free. This gives rise to a *nonstationary multi-class finite-server queuing system*. Henceforth, without loss of generality, we will concentrate our discussion on a single multi-class queue associated with one resource group.

Our aim is to compute the expected number of customers (i.e., units of LRUs) in the queue of each class (i.e., LRU type) at any time. To our knowledge, there is no known analytical method on how to compute this measure exactly. In the following, we propose a new computationally efficient approximation method. We assume that the inter-arrival and service time follow exponential distributions, respectively.

In Rothkopf and Oren (1979), a method was proposed that gives good approximations for nonstationary single-class queuing systems. In Alfredsson (1997) and Díaz and Fu (1997), the authors view multiple classes of customers as a single class by using cumulative arrival rate and mean service rate. Our proposed strategy is essentially to first *merge* multiple classes into a single class as Alfredsson (1997) and Díaz and Fu (1997), and employ the method in Rothkopf and Oren (1979) to estimate the expected number of customers in the system of all classes. We then calculate the expected number of customers of each class via *disaggregation*. Details are as follows.

Before presenting our method, we introduce some further notations. Notice that in our attempt to follow queuing terminology in this section, we will unambiguously equate the index c (customer class) with index k (LRU type).

s: Number of servers

$\lambda_c(t)$: Arrival rate of class c at time t (equivalent to $\lambda_{0k}(t)$ for LRU k)

μ_c: Service rate of class c (equivalent to TAT_k for each LRU k)

$\lambda(t)$: Cumulative arrival rate at time t

$\mu(t)$: Mean service rate at time t

$N_c(t)$: Expected number of class c customers in the system at time t (which is the target to be computed)

$N(t)$: Expected number of customers of all classes in the system at time t

$Q_c(t)$: Expected number of class c customers waiting in queue at time t

$Q(t)$: Expected number of customers of all classes waiting in queue at time t

$R_c(t)$: Expected number of class c customers in service (i.e., being repaired) at time t

$U(t)$: Server utilisation (i.e., expected number of customers of all classes in service) at time t.

The cumulative arrival rate at time t is the summation of arrival rates of all classes at time t, given by

$$\lambda(t) = \sum_c \lambda_c(t) \tag{12}$$

and the mean service rate of all classes at time t is given by

$$\mu(t) = \frac{\lambda(t)}{\sum_c (\lambda_c(t)/\mu_c)}. \tag{13}$$

From Matta et al. (1995) and Rothkopf and Oren (1979), we know how to compute the expected number of customers of all classes in the queuing system at time t. The algorithm is given in Rothkopf and Oren (1979), starting from $t = 0$ when the system is empty. After computing $N(t)$, we calculate the time-dependent expected number of each class customers in the system. We have for all t:

$$N(t) = \sum_c N_c(t). \tag{14}$$

The key is to separate $N(t)$ into disjoint parts, one for each class.

We know from Alfredsson (1997) and Díaz and Fu (1997) that the expected waiting time for a server is equal for all customers under steady state. Let W be the expected waiting time for a server under steady state. We have:

$$N_c = \lambda_c W + \lambda_c/\mu_c \tag{15}$$

for stationary models. If we divide the expected number of customers in the system into two parts: one in queue Q_c, and the other in service R_c, then, from equation (15), we see that $Q_c = \lambda_c W$, which is proportional to the arrival rate of class c, and $R_c = \lambda_c/\mu_c$, which is proportional to the server utilisation for class c.

Under the *time-dependent* case, since the in-rate is dependent on the arrival rate $\lambda(t)$ and out-rate is dependent on both the arrival rate $\lambda(t)$ and service rate $\mu(t)$, we will divide the expected number of customers of all classes in the system at time t $N(t)$ into two parts: one is that in queue $Q(t)$ and the other is that in service i.e., $U(t)$. Hence,

$$N(t) = Q(t) + U(t) \tag{16}$$

$$N_c(t) = Q_c(t) + R_c(t). \tag{17}$$

Since the expected number of each class customers in queue and in service is proportional to the respective arrival rate and server utilisation, i.e., $Q_c(t)$ is proportional to $\lambda_c(t)$ and $R_c(t)$ is proportional to $\lambda_c(t)/\mu_c$, we have:

$$Q_c(t) = \frac{\lambda_c(t)}{\lambda(t)} Q(t) = \frac{\lambda_c(t)}{\lambda(t)} (N(t) - U(t)) \tag{18}$$

$$R_c(t) = \frac{\lambda_c(t)/\mu_c}{\lambda(t)/\mu(t)} U(t). \tag{19}$$

6 Optimisation

In this section, we present our optimisation algorithm to solve the corrective maintenance problem. Since the system performance EBO varies with time, we are in fact dealing with a time-dependent optimisation problem. Most literature deals with either steady-state results or optimisation with respect to a certain time point of interest. For example, in Isaacson and Boren (1988) and Slay et al. (1996), the time at the end of peak demand rate/utilisation rate, which is viewed as the 'worst' or most demanding day is selected as the point of interest since it is believed that the allocation of spares sufficient to support that point is also adequate to maintain the target throughout the life cycle.

Unfortunately, our problem is not as straightforward. Since we are under finite repair resources, not all failures can be repaired at once. Therefore, the EBO at the end of peak utilisation rate may *not* be the worst case due to the unavailability of repair resources. It is also not surprising that with finite repair resources, where the repair time is long and the demand rate is high, the number of failures in the repair facility pipeline will build up over time. In other words, if we myopically optimise with respect to the time point at the end of peak utilisation rate, the spares and repair resources, which are sufficient to support that point, may not be sufficient to maintain the system performance throughout the operating horizon. This implies that instead of choosing the end of the peak utilisation as the gauge for the worst case EBO, we should indeed be examining EBO at various time points in the horizon and locate the time point when the worst case occurs.

Since time t is a continuous variable, ranging from 0 to the operating horizon T, we first discretise the time horizon into N periods, which are indexed by $n = 1, \ldots, N$, so that

- the utilisation rate within each period is constant
- the length of each period is small enough so we do not miss out on the worst case.

We define t_n to be the time at the end of period n, so $maxEBO = \max_{n \in \{1,\ldots,N\}} EBO(t_n)$.

6.1 Significance levels and resource groups

In OPUS9 (1992) and OPUS10 (1998), the concept of *significance level* was introduced. The significance level describes the importance of different stock positions, where a stock position contains an item and its location. The more important the stock position, the higher the significance level is. In our problem, the stock positions of spares at bases are the most important because they have direct effects on EBO, whereas the spares at the

depot are of next importance because they determine the resupply delay for LRUs. The repair resources at the depot are the least important stock position, only influencing the repair pipeline of LRUs at the depot. Consequently, the significance level of repair resources at the depot is set to be 0, that of spares at the depot is 1, and that of spares at the bases is 2. Since the computation of EBO at a certain level only requires the information on stock positions whose significance levels are less or equal, optimisation can be carried out level-by-level from the lowest-significance level 0 until the highest level.

Under the assumption that each failed LRU requires only one repair resource (see Assumption (5) in Section 3), different LRUs can be partitioned into disjoint *resource groups* in such a way that each group contains the common repair resources and LRUs requiring them. In doing so, our strategy is to decompose the entire problem into independent sub-problems for each resource group, and subsequently combine them using marginal allocation.

6.2 Optimisation algorithm for generating Cost/Effectiveness (C/E) curve

We now present our optimisation algorithm to generate the C/E curve. Our approach follows the optimisation algorithm proposed in Alfredsson (1997) (which was based on marginal analysis in Sherbrooke, 1992), and extends it by the time dimension. Marginal analysis is essentially a myopic method where one unit of the item with maximum marginal utility is increased at each step. The basic idea is to apply marginal analysis from the lowest to the highest significant levels, and then merge the C/E curves via convexification, in the same way as Alfredsson (1997) and Sherbrooke (1992).

The algorithm proceeds in three stages. First divide all LRUs according to disjoint *resource groups*. Second, within each resource group, perform optimisation (i.e., generate a C/E curve for this resource group) as follows:

1 Apply marginal analysis to generate the C/E curve for significance level 0. (This curve contains optimal allocations of repair resources at the depot)

2 For each repair resource allocation in (1), generate a C/E curve by doing the following:

   2a For each LRU within the resource group, find the time point t^* when the worst case depot EBO occurs (i.e., $EBO_{0k}(t^*) = \max_{n\in\{1,\ldots,N\}} EBO_{0k}(t_n)$) and obtain spares allocation with respect to t^* as follows:

      i At time t^*, apply marginal analysis to generate the C/E curve for significance level 1.

      ii For each point on the step (1) C/E curve, apply marginal analysis at time t^* to generate a C/E curve for significance level 2.

   2b Apply marginal analysis to combine the curves obtained in step 2a to generate a C/E curve for this resource allocation.

3 Merge all C/E curves generated in step (2) to form the C/E curve for this resource group. (Note that this merging can be performed efficiently by any standard computational geometry algorithm, such as the Graham scan, that finds a convex hull for a given set of points on the plane).[6]

Third and finally, we apply marginal analysis to combine the C/E curves of all repair resource groups to generate the final C/E curve. Details of our algorithm are found in Appendix A.

7 Experimental results

We consider a number of test cases extracted from real-life scenarios of a military organisation. There are 40 identical military systems deployed at the bases, each of which has 46 LRUs with very different failure rates and repair times. For example, repair time of some LRUs is 168 h while that of others is 1556.6 h while *MTBF* ranges from 2000 h to 10,309 h. We will benchmark our results against a specialised simulation tool (that behaves identically as SPAR). For the purpose of benchmarking, we will use the time-dependent availability (*Ao*) as the system performance. Note again that conversion from *EBO* into *Ao* in a time-dependent setting has been presented in Lau et al. (2006). For simplicity, we shall term our proposed approach as 'approximation' or 'App.' in the following tables, in contrast with 'simulation' or 'Sim.' in the following tables (for the simulation tool).

The large number of experiments and their extensive results do not allow us to present all the details in this paper. We will only present the key performance characteristics.

7.1 Evaluation of system performance

To assess the accuracy of our proposed approximation over a given spare and repair resource allocation, we compare results from a simulation run. The simulation experiments were developed on the Extend (version 6) software, in which we use the random number generator provided to generate the time between arrivals and service time based on exponential distribution. The simulation result is based on the average of the performance measures obtained by running 1000 replications (which enabled us to achieve results within 95% confidence interval).

First, we assess the quality of our approximate method for the nonstationary multi-class finite-server queue presented in Section 5. We ran a large number of cases among which the result of one case is provided here (due to space constraint). In this case, there are three servers and two classes of customers whose service rates are $\mu_1 = 0.2$, $\mu_2 = 0.25$. The demand rates are time varying, which are given as follows: λ_1 = 1/3(0–48 h), 1/12(48–120 h), 1/6(120–168 h); λ_2 = 1/4(0–48 h), 1/16(48–120 h), 1/8(120–168 h). The results are shown in Figure 1. From that we can see that our analytical model agrees well with simulation results.

Next, we experiment on instances arising from a real application. In the following problem instance, the utilisation rates are given as follows: 0.0486 (0–1440 h), 0.0833 (1440–2160 h), 0.5833 (2160–2520 h), 0.3333 (2520–3240 h), 0.5 (3240–3600 h) and 0.25 (3600–4320 h). We assume that there is one repair resource type, which can repair all LRUs. In this instance, the MTBF values are large, ranging from 2000 h to 10,309 h, implying that very few demands are entering the system, which causes the number of items in queue and service to be very small. We randomly choose five LRUs from the data set and ran 1000 simulation replications to generate both the average number of items in queue and service as well as their 95% confidence intervals. Again, Table 2

shows that our analytical approximates are all within 95% confidence intervals. We then performed experiments to evaluate the system performance. The results are shown in Figure 2. From Figure 2, we observe that our approximation matches simulation very well with low error. The highest error is around 3.5% and the error for availability under worst case is around 0.04%. Next, we demonstrate that stationary approximations fail miserably in a time-dependent setting. For this purpose, we first approximate the nonstationary demand with a stationary demand with the average utilisation rate of 0.217575 (0–4320 h). We then compute *Ao* using our proposed model but with a standard Poisson demand instead. Figure 2 shows the gap in terms of actual system behaviour between the stationary (Stat.) and nonstationary (App.) approximations. Observe also that the trends of two curves are totally different, the highest error is around 57% and the error for availability under worst case is more than 7.5%. Figure 3 shows the system behaviours of high availability using simulation (Sim.), nonstationary (App.) approximation and stationary (Stat.) given more spares.

Figure 1 Expected number of customers of each class in the nonstationary system

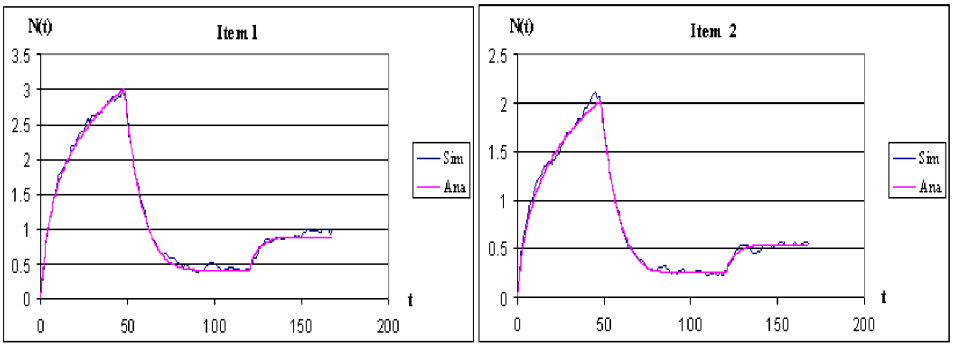


Table 2 Approximation vs. simulation by 95% confidence interval

Time	*Class c*	*Sim.*	*LB*	*UB*	*App.*	*In 95% interval*
48 h	Class 1	0.111	0.090370	0.131630	0.125607	True
	Class 2	0.065	0.048737	0.081263	0.070686	True
	Class 3	0.051	0.036805	0.065195	0.061391	True
	Class 4	0.162	0.135130	0.188870	0.143537	True
	Class 5	0.043	0.030118	0.055882	0.035318	True
120 h	Class 1	0.196	0.163922	0.228078	0.172506	True
	Class 2	0.103	0.082397	0.123603	0.100819	True
	Class 3	0.076	0.058218	0.093782	0.086738	True
	Class 4	0.227	0.194929	0.259071	0.214665	True
	Class 5	0.045	0.031560	0.058440	0.051765	True
168 h	Class 1	0.203	0.170199	0.235801	0.185531	True
	Class 2	0.092	0.070922	0.113078	0.109611	True
	Class 3	0.072	0.055038	0.088962	0.084052	True
	Class 4	0.266	0.230418	0.301582	0.236403	True
	Class 5	0.059	0.043127	0.074873	0.056702	True

Next, we like to quantifiably verify the statement in Sleptchenko et al. (2002), "the impact of finite capacity increases with repair shop utilisation and decreases with the number of servers". That is, we like to investigate the effects of finite capacity on availability. Unlike the queuing model under steady state in which the utilisation rate must be less than 1, which otherwise cannot go into steady state, the utilisation rate under a time-dependent scenario can be greater than 1. Since there are 40 military systems deployed at the base, the utilisation rate can be very large. Under the given stock allocation (provided by an industry partner), Table 3 shows that our solution achieves more than 95% availability under infinite repair capacity (see last column). From Table 3, we can verify the impact of finite capacity stated in Sleptchenko et al. (2002): that the deviation (gap) in availability decreases with capacity and increases with utilisation rate.

Figure 2 *Ao* by simulation, approximation and stationary with low *Ao*

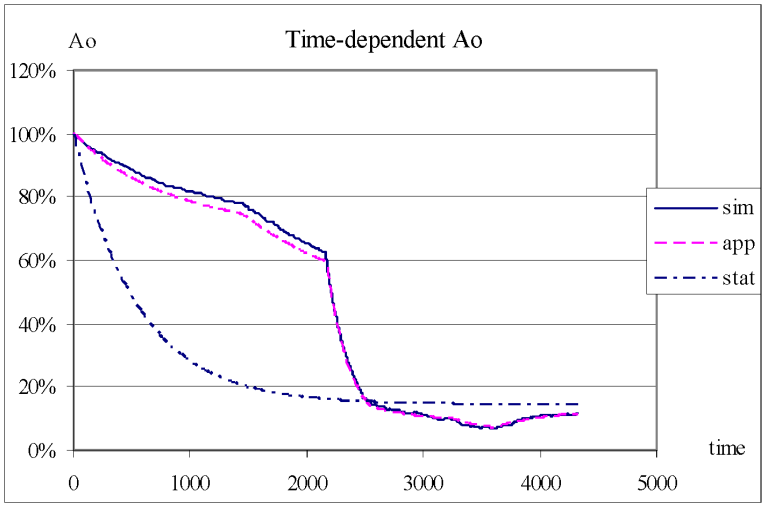


Figure 3 *Ao* by simulation, approximation and stationary with high *Ao*

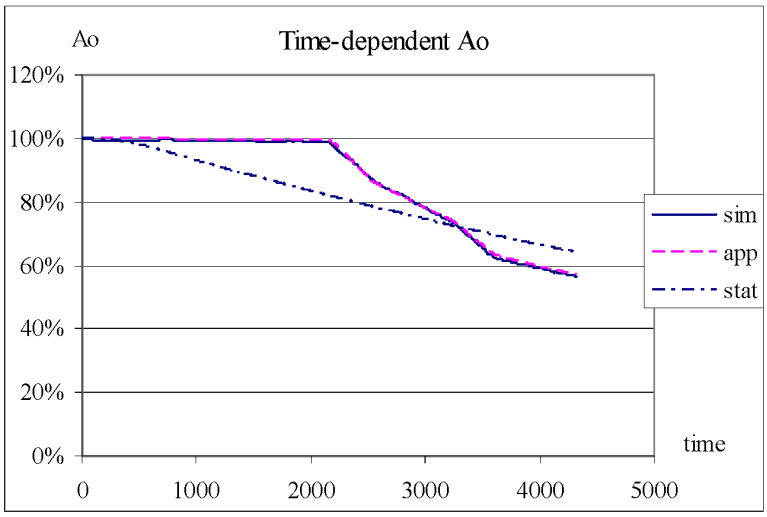


Table 3 Impact of finite capacity (Capacity and UR of repair resources)

Capacity	*UR*	*App.* (%)	*Sim.* (%)	∞ *Capacity* (%)
1	60.18	35.82	36.10	96.99
5	12.04	40.91	41.39	96.99
10	6.02	46.98	46.80	96.99
25	2.41	63.93	63.64	96.99
40	1.50	78.47	78.94	96.99
65	0.93	94.86	94.47	96.99
80	0.75	96.93	96.73	96.99

Next, we examine the impact of finite capacity as the change of spare allocation. The results are shown in Table 4, where columns 2–4 show the availabilities achieved under repair capacity (rc) of 10 units, 25 units, and infinite number of units under different spares cost, respectively. $Err_{3,4}$ is the deviation between columns 3 and 4. From Table 4 $Err_{3,4}$ column, we observe interestingly that the impact of finite capacity will increase with increasing spares first and then decrease after a certain level. This can be explained as follows. In the case of infinite repair capacity, all increased spares will improve availability. Where there are finite capacities, however, increased spares will barely offset the effect of finite capacities at the beginning; but after a certain threshold, the offset takes effect – since all newly increased spares will replace failures immediately even though there are finite capacities. Hence, the impact of finite capacity increases with the spare and then decreases after a certain level. This phenomenon takes place not only between finite and infinite repair capacity, but also between fewer and greater capacity (See $Err_{2,3}$).

Table 4 Impact of finite capacity as spares

Cost	*App.* (*rc = 10*) (%)	*App.* (*rc = 25*) (%)	∞ *Capacity* (%)	$Err_{3,4}$ (%)	$Err_{2,3}$ (%)
0	14.41	32.84	36.59	3.75	18.43
6538	14.58	35.17	47.14	11.97	20.59
52564	16.63	37.73	66.26	28.53	21.10
78080	20.13	40.81	84.45	43.64	20.68
115761	28.62	47.51	87.33	39.82	18.89
266532	46.98	63.93	96.99	33.06	16.95
646447	79.40	94.99	99.99	5.00	15.59

Tables 5 and 6 provide an insight into the impact of nonstationary demand approximation. Since we are concerned with the system performance under the worst case, the availabilities listed in Tables 5 and 6 are the worst availabilities over the operating horizon. In Table 5, the worst availabilities obtained by simulation (Sim.), our approximation method (App.) and the stationary approximation under varying repair resource capacities (Stat.) are listed, and we compute the errors from simulation, respectively. Similarly, Table 6 presents comparison results under varying spare allocations with different costs while fixing resource capacity at 25. Table 5 shows that

error gets larger as the number of resources increases while Table 6 shows that the error gets smaller as more spares are bought. This can be explained intuitively as follows. When the repair capacity increases at the beginning, due to heavy utilisation in nonstationary scenario, the size of pipeline may not be improved as well as that under smooth utilisation rate, so the error gets larger. Until both the queuing systems converge towards steady state, that is, the effect of capacity tends towards saturation, the deviation gets smaller and tends to be steady. That explains why the error decreases after 16.44% and around 14.2% eventually (see Table 5). On the other hand, although the deviation under worst case gets smaller as more spares are bought, the stationary approach still cannot approximate the system behaviour effectively, especially during transition periods (See Figure 3).

To illustrate the phenomenon that the system performance at the end of peak utilisation rate may not be the worst case, we use three simplified test cases with only one LRU whose *MTBF* is 10,000 h. Table 7 gives a summary of the results. $TAT = 1556.5$ h in the first case and $TAT = 168$ h in the second case. The operating horizon T is 4320 h and the utilisation rates are given as above. In the third case, both $TAT = 168$ and $T = 168$ and $UR = 1$(0–48 h), 1/13(48–120 h), 0.5(120–168 h). In Table 7, the column *EBO*(t) shows the *EBO* at the end of peak utilisation rate and the corresponding time point at which this EBO occurs, while the column *maxEBO*(t) is the worst *EBO* and the corresponding time point. From Table 7, we can see for test cases 1 and 2, the worst system performance takes place at the time far beyond the end of peak utilisation rate when repair capacity is small and only gets closer as the capacity increases. Comparing case 1 with 2, we can see the worst EBO occurs at the time, which gets closer to the end of peak utilisation rate more quickly when *TAT* is small. This can be explained as: when capacity is small, there are too few resources to repair the failed LRUs in time so the pipeline gets larger with time, whereas with enough capacities ailed units can be repaired almost at once. On the other hand, small *TAT* means repair can be finished within a short duration so that this item in the pipeline will not be counted over time, while large *TAT* means it will always be kept in the pipeline at least for a long repair time. Comparing test cases 2 and 3, we can see when the operating horizon is short, even shorter than *TAT*, the worst EBO will occur at time far beyond the end of peak utilisation rate.

Table 5 Impact of nonstationary approximation as capacity

Capacity	*Sim.* (%)	*App.* (%)	*Stat.* (%)	*Err_App* (%)	*Err_Stat* (%)
0	0	0.03	0.04	0.03	0.04
5	3.60	3.59	7.23	–0.01	3.63
8	5.60	5.66	11.54	0.06	5.94
10	6.80	7.04	14.41	0.24	7.61
12	7.80	8.41	17.27	0.61	9.47
20	14.00	13.79	27.65	–0.21	13.65
22	14.60	15.04	29.89	0.44	15.29
25	16.40	16.79	32.84	0.39	16.44
28	19.60	18.31	34.96	–1.29	15.36
30	21.30	19.16	35.78	–2.14	14.48
100	22.40	21.06	36.53	–1.34	14.13
Infinite	22.40	22.54	36.59	0.14	14.19

Table 6 Impact of nonstationary approximation as spares cost

Cost	*Sim.* (%)	*App.* (%)	*Stat.* (%)	*Err_App* (%)	*Err_Stat* (%)
0	16.40	16.79	32.84	0.39	16.44
6538	17.90	18.30	35.17	0.40	17.27
52564	24.20	23.90	37.73	–0.30	13.53
78080	30.60	30.42	40.81	–0.18	10.21
115761	39.50	38.38	47.51	–1.12	8.01
266532	56.50	56.80	63.93	0.30	7.43
646447	91.70	91.50	94.99	–0.20	3.29

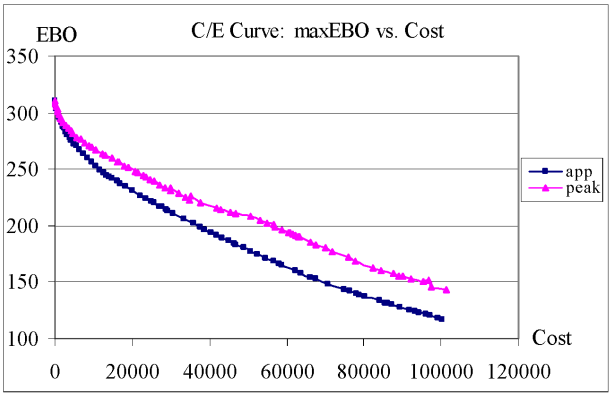
Table 7 Peak demand EBO vs. worst case EBO

Case	*Capacity*	*EBO* (*t*)	*maxEBO* (*t*)	*Error* (%)
TAT = 1556.6 (T = 4320)	0	1.360 (2520)	3.760 (4320)	63.83
	1	1.055 (2520)	2.532 (4320)	58.34
	3	0.996 (2520)	1.793 (3600)	44.48
	5	0.994 (2520)	1.752 (3600)	43.27
	8	0.994 (2520)	1.749 (3600)	43.19
TAT = 168 (T = 4320)	0	1.360 (2520)	3.760 (4320)	63.83
	1	0.413 (2520)	0.423 (3600)	2.50
	3	0.353 (2520)	0.353 (2520)	0.00
	5	0.353 (2520)	0.353 (2520)	0.00
TAT = 168 (T = 168)	0	0.192 (48)	0.310 (168)	38.10
	1	0.169 (48)	0.187 (168)	9.88
	3	0.167 (48)	0.179 (168)	6.53

7.2 Optimisation

The above section illustrates that our proposed approximation approach leads to a fairly accurate system performance. In this section, we validate the effectiveness of our proposed optimisation algorithm. We apply it to a real-world large test case where each military system has more than 50 LRUs. The results are shown in Figure 4. We compare our approach with the standard marginal analysis that selects the time point at the end of peak utilisation rate as the worst-case time point (t^*). The EBO in Figure 4 is the worst case over the whole operating horizon. From Figure 4, we find that the results by selecting peak utilisation rate are not optimal and not even convex when compared with ours. As discussed in Section 6, this is because the worst EBO is not at the end of peak utilisation rate due to finite repair resource and long repair time.

Figure 4 C/E curves with operating horizon vs. only the end of peak UR



8 Conclusion and future research

Our work involves an innovative hybridisation of analytical modelling, optimisation, queuing theory, and simulation to solve a 2-echelon repairable item inventory problem under nonstationary Poisson demands and finite repair facilities.

In our work, the assumption that each LRU requires exactly one repair resource is crucial in reducing the complexity of the problem. What is challenging for future work is to develop analytical models that relax this assumption. Another interesting research is to adopt priority queuing model for repair by relaxing the assumption of FCFS discipline. We also assume that the utilisation rates are the same across all bases so that the demand rates vary synchronously, so another challenging aspect of work is the development of a model where utilisation rates vary with different systems at different bases at different times. Finally, the spare allocation is fixed over operating horizon once it is determined in our model. A natural extension of future work is to allow reallocation mission by mission over a rolling horizon based on existing system performances in previous periods.

Acknowledgement

A preliminary version of this paper appeared in *Proceedings of the National Meeting* of the Decision Sciences Institute, 1901–1908, Boston, USA, November 2004.

References

Alfredsson, P. (1997) 'Optimization of multi-echelon repairable item inventory systems with simultaneous location of repair facilities', *European Journal of Operational Research*, Vol. 99, pp.584–595.

Alfredsson, P. (1999) *OPRAL – A Model for Optimum Resource Allocation*, Systecon AB, SE-102 45 Stockholm, Sweden.

Carrillo, M.J. (1991) 'Extensions of Palm's theorem: a review', *Management Science*, Vol. 37, pp.739–744.

Cohen, M.A., Zheng, Y.S. and Agrawal, V. (1997) 'Service parts logistics: a benchmark analysis', *IIE Transactions*, Vol. 29, pp.627–639.

Díaz, A. and Fu, M.C. (1997) 'Models for multi-echelon repairable item inventory systems with limited repair capacity', *European Journal of Operational Research*, Vol. 97, pp.480–492.

Graves, S. (1985) 'A multi-echelon inventory model for a repairable item with one-for-one replenishment', *Management Science*, Vol. 31, No. 10, pp.1247–1256.

Green, L. and Kolesar, P. (1991) 'The pointwise stationary approximation for queues with nonstationary arrivals', *Management Sci.*, Vol. 37, No. 1, January, pp.84–97.

Green, L. and Kolesar, P. (1997) 'The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates', *Management Science*, Vol. 43, No. 1, January, pp.80–87.

Green, L., Kolesar, P. and Svoronos, A. (1991) 'Some effects of nonstationarity on multiserver Markovian queueing systems', *Operations Research*, Vol. 39, pp.502–511.

Isaacson, K.E. and Boren, P. (1988) 'Dyna-METRIC Version 5: a capability assessment model including constrained repair and management adaptations', *The RAND Corporation*, Technical Report R-3612-AF, Santa Monica, CA.

Jung, W. (1993) 'Recoverable inventory systems with time-varying demand', *Production and Inventory Management Journal*, Vol. 34, No. 1, pp.77–81.

Kim, J.S., Shin, K.C. and Park, S.K. (2000) 'An optimal algorithm for repairable-item inventory system with depot spares', *Journal of the Operational Research Society*, Vol. 51, pp.350–357.

Lau, H.C., Song, H., See, C.T. and Cheng, S.Y. (2006) 'Evaluation of time-varying availability in multi-echelon spare parts systems with passivation', *European Journal of Operational Research*, Vol. 170, No. 1, pp.91–105.

Matta, I. and Shankar, A.U. (1995) 'Z-iteration: a simple method for throughput estimation in time-dependent multi-class system', *Proc. ACM SIGMETRICS/PERFORMANCE'95*, *ACM SIGMETRICS Performance Evaluation Review*, Vol. 23, No. 1, May, Ottawa, pp.126–135.

Nelson, B.L. (1995) *Stochastic Modeling: Analysis and Simulation*, Dover Publ. Inc., ISBN 0-486-42569-X.

OPUS10 (1998) *User's Reference – Logistics Support and Spares Optimization*, Version 3, Systecon AB, May.

OPUS9 (1992) *Version 1.6 Users Guide*, Systecon AB, January.

Perlman, Y., Mehrez, A. and Kaspi, M. (2001) 'Setting expediting repair policy in a multi-echelon repairable-item inventory system with limited repair capacity', *Journal of the Operational Research Society*, Vol. 52, pp.198–209.

Pyke, D.F. (1990) 'Priority repair and dispatch policies for repairable-item logistics systems', *Naval Research Logistics*, Vol. 37, pp.1–30.

Rothkopf, M.H. and Oren, S.S. (1979) 'A closure approximation for the nonstationary M/M/s queue', *Management Science*, Vol. 25, pp.522–534.

Sherbrooke, C.C. (1968) 'METRIC: a multi-echelon technique for recoverable item control', *Operations Research*, Vol. 16, No. 2, pp.122–141.

Sherbrooke, C.C. (1986) 'VARI-METRIC: improved approximation for multi-indenture, multi-echelon availability models', *Operations Research*, Vol. 34, No. 2, pp.311–319.

Sherbrooke, C.C. (1992) *Optimal Inventory Modeling of System: Multi-Echelon Techniques*, John Wiley & Sons, New York.

Slay, F.M., Bachman, T.C., Kline, R.C., O'Malley, T.J., Eichorn, F.L. and King, R.M. (1996) *Optimizing Spares Support: The Aircraft Sustainability Model*, October, Technical Report, Logistics Management Institute, McLean, Virginia.

Sleptchenko, A., van der Heijden, M.C. and van Harten, A. (2002) 'Effects of finite repair capacity in multi-echelon, multi-indenture service part supply systems', *International Journal of Production Economics*, Vol. 79, pp.209–230.

Sleptchenko, A., van der Heijden, M.C. and van Harten, A. (2003) 'Trade-off between inventory and repair capacity in spare part networks', *Journal of the Operational Research Society*, Vol. 54, pp.263–272.

Zijm, W.H. and Avsar, Z.M. (2003) 'Capacitated two-indenture models for repairable item systems', *International Journal of Production Economics*, Vols. 81–82, No. C, pp.573–588.

Notes

[1]In this paper, we use the word '(military) system' to denote equipment such as an aircraft, ship or tank. This is not to be confused with the term '(inventory) system' or '(queuing) system' also used in the paper. The context should be clear in each occurrence of the term.

[2]In this paper, a nonstationary Poisson process refers to a Poisson process, which has a time-dependent arrival rate (see, for example Nelson, 1995).

[3]For the sake of simplicity, we assume that repair resources comprise only of manpower.

[4]This was proven in an internal technical report, which is available upon request.

[5]Like Perlman et al. (2001), we assume $Nsys_j$ is large enough that the failure may depend upon the required number of systems, but did not depend on the actual number of working systems. Furthermore, in the Air Force environment, this assumption holds because the fleet must maintain the same number of flight missions regardless of the actual number of working aircrafts.

[6]For an animation of this algorithm, see for example www.cs.princeton.edu/~ah/alg_anim/version0/Graham.html

Website

SPAR Website, http://www.clockwork-group.com

Appendix A: Optimisation Algorithm Pseudo-Code

Let each resource group be denoted by RG_g, $g = 1, \ldots, G$. Hence, according to equations (1) and (8), we may rewrite system performance and investment cost as a linear sum of the respective resource groups:

$$EBO(t) = \sum_{g=1}^{G} EBO_g(t), \quad LSC = \sum_{g=1}^{G} LSC_g$$

where $EBO_g(t) = \sum_{k \in RGg} \sum_{j=1}^{J} EBO_{jk}(t)$ (i.e., total EBO for those LRUs within resource group g) and $LSC_g = Cr_g r_g + \sum_{k \in RGg} Cs_k \left(\sum_{j=0}^{J} s_{jk} \right)$ (i.e., total life support cost for those LRUs within resource group g). Let $EBO_g^L (L = 0, 1, 2)$ denote the EBO associated with significant levels 0, 1, and 2, respectively. The pseudo-code is given as follows:

1 Divide LRUs into resource groups, and for each resource group RG_g, perform steps (2)–(5).

2 Initialise *EBO* and *LSC* for repair resource at significance level 0 defined by $EBO_g^0 = \max_{n\in\{1,\ldots,N\}} EBO_g^0(t_n) = \Sigma_{k\in RGg} EBO_{0k}(t_n), LSC_g^0 = 0$.

3 Until ($LSC_g^0 >$ Budget or EBO_g^0 is not improved), increase r_g by 1 and update EBO_g^0, LSC_g^0, which generates a C/E curve of significance level 0.

4 For each repair resource allocation on the above C/E curve do the following two steps:

4(a) For each LRU k ($k \in RG_g$).

(i) Find the time point t^* such that $EBO_{0k}(t^*) = \max_{n\in\{1,\ldots,N\}} EBO_{0k}(t_n)$.

(ii) Initialise EBO and LSC at significance level 1 defined by $EBO_{gk}^1(t^*) = EBO_{0k}(t^*)$ and $LSC_{gk}^1 = LSC_g^0 + Cs_k s_{0k} (k \in RG_g)$.

(iii) Until ($LSC_{gk}^1 >$ Budget or $EBO_{gk}^1(t^*)$ is not improved), increase s_{0k} by 1 and update $EBO_{gk}^1(t^*), LSC_{gk}^1$. This will generate a C/E curve of significance level 1.

(iv) Initialise *EBO* and *LSC* at significance level 2 defined by $EBO_{gk}^2(t^*) = \sum_{j=1}^{J} EBO_{jk}(t^*)$ and $LSC_{gk}^2 = LSC_{gk}^1 + Cs_k \sum_{j=1}^{J} s_{jk}, (k \in RG_g)$.

(v) Until ($LSC_{gk}^2 >$ Budget or $EBO_{gk}^2(t^*)$ is not improved), increase s_{j^*k} by 1 whose marginal utility $MU_{j^*k} = \max_{j\in\{1,2,\ldots,J\}} MU_{jk}$ where $MU_{jk} = |\Delta EBO_{jk}(t^*)|/Cs_k$. This will generate a C/E curve of significance level 2.

4(b) Apply marginal analysis to combine all level-2 C/E curves generated above to generate a C/E curve for the resource allocation.

5 Merge all C/E curves associated with all resource allocations to generate a C/E curve for the resource group.

6 Combine all C/E curves for all resource groups to generate a final C/E curve for the entire problem.