ELSEVIER

Int. J. Production Economics 79 (2002) 209–230

international journal of production economics

www.elsevier.com/locate/dsw

Effects of finite repair capacity in multi-echelon, multi-indenture service part supply systems

A. Sleptchenko*, M.C. van der Heijden, A. van Harten

Faculty of Technology and Management Operation Methods and Systemtheory, University of Twente, P.O. Box 217, 7500 AE Enschede, Netherlands

Received 1 June 2001; accepted 26 March 2002

Abstract

In this paper, we consider multi-echelon, multi-indenture supply systems for repairable service parts with finite repair capacity. We show that the commonly used assumption of infinite capacity may seriously affect system performance and stock allocation decisions if the repair shop utilisation is relatively high. Both for the case of item-dedicated and shared repair shops, we modify the well-known VARI-METRIC method to allocate service part stocks in the network. The repair shops are modelled by (single or multi-class) multi-server queuing systems. We validate our procedure by comparison to results from discrete event simulation. This comparison shows that the accuracy of the technique presented in this article is on average more than five times as close to simulated values as the classical VARI-METRIC technique. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Spare parts; Multi-echelon inventory; Multichannel queues

1. Introduction

Technically advanced systems play an ever more important role in society. As a consequence, the availability of such systems may strongly affect daily operations. This applies to, e.g. heavily automated production processes, computer systems, medical equipment, and military systems. Downtime of critical equipment may have serious consequences, e.g. in terms of loss of production, quality reduction in health care or ineffective military missions. Various measures can be taken to reduce the amount of system downtime, such as system redundancy, appropriate preventive maintenance and effective corrective maintenance. Especially with respect to the latter, fast supply of the service parts required is essential. Here we define service parts as all parts that are used to maintain systems, both spare parts to replace failed parts and diagnostic items that are used to analyse the system performance and to find failure causes.

The importance of service parts management has increased in the past decades. One reason is the fact that system availability and high quality after sales service have become important criteria when selecting suppliers of technically advanced systems. A second reason is the increasing value of service part inventory investment. A survey by Cohen et al. [1] reports that service parts

*Corresponding author.

E-mail address: a.sleptchenko@sms.utwente.nl (A. Sleptchenko).

0925-5273/02/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0925-5273(02)00155-X

inventories equal 8.75% of the value of product sales in their sample, being over $23 million inventory investment on average. In this survey, the following characteristics and trends in service delivery organisations are observed:

1. a large and geographically dispersed installed base;
2. a large number of service parts to be stocked, varying between 2500 and 300,000 in the sample;
3. increasing costs of service parts due to increasing complexity and modularity, in the sample being $270 on average with exceptions up to several hundreds thousands of dollars;
4. a high part obsolescence rate caused by short product life cycles;
5. a large and increasing fraction of slow moving, caused by increased system customisation, and design improvement; the average inventory turnover in the sample equals 0.87 parts per year, hence there are many parts having a demand rate less than once per year.

As a consequence of the increasing costs of spare parts (third characteristic above), it is worthwhile to consider repair rather than scrap and replace. In the survey by Cohen et al. [1], it appears that in the sample 27.2% of the parts are *repairable*, i.e. parts for which repair is technically possible and economically profitable. Although this suggests that consumables are more important, one should keep in mind that repairables are generally more expensive, so the share of repairables in total service part investment is probably considerably higher.

Service parts are often supplied via a multi-echelon distribution network, i.e. a hierarchical network of stocking locations through which service parts are supplied to the customer's site. A reason to have a multi-echelon structure is the need for both local stocks close to the customer's sites in order to achieve fast supply and the need for stock centralisation to reduce holding costs. Cohen et al. [1] report that three-echelon networks are prevalent in their sample followed by two-echelon systems. Four-echelon networks occur in practice as well. There is a trend however to reduce the number of echelons and the number of locations per echelon in order to reduce fixed warehousing costs and service parts obsolescence costs. This striving for lean and efficient service part networks is facilitated by, e.g. stocking essential parts at the customer sites and using possibilities for fast emergency transportation.

All these characteristics cause that service parts management is an increasingly important, yet complex task. A key challenge is to attain high availability of the installed base at low service costs. These service costs include costs for stock holding, warehousing, transportation, service engineers, repair shops and overhead. Effective and efficient spare part management means that several design choices have to be made (e.g. network structure) and a suitable logistical control structure has to be developed. For a nice overview of the key issues for the logistical control of service part supply systems, we refer to the framework of Verrijdt [2].

One way to influence the relation between customer service (high availability of the installed base) and costs of the service part supply system is by appropriate stock allocation. That is, decisions have to be taken about which parts to stock at which locations in the network in which amounts. This well-known stock allocation problem for service networks is different from traditional inventory models, because the relevant performance measure is availability of the installed base rather than part fill rates. A specific availability level can be attained by several combinations of part fill rates. One option is to use an equal fill rate for each part such that the target availability is attained. A more clever option is to focus on high fill rates for cheap, slow movers, so that relatively low fill rates (and hence low stock levels) of expensive, fast movers are sufficient. In this way, the same system availability can be obtained at lower costs.

Many models for these kinds of stock allocation problem have been developed in the past decades. Already in the 60s, Sherbrooke [3] developed the famous METRIC model for repairable item inventory control. This model has been the basis for a lot of additional research later on. An overview of the most important models based on the METRIC approach is given in Sherbrooke [4].

One of the key assumptions made by Sherbrooke is that all repair shops in the network have infinite capacity. Although this is generally not true in practice, this assumption may be justified by some arguments. Firstly, in some situations the repair of service parts is just one of the activities of the repairmen, e.g. next to preventive maintenance. If service part repair gets highest priority, the effect of finite capacity can be negligible. However, this is only true if service part repair is not a major task for each repairman. Besides, preventive maintenance can have high priority as well, e.g. when preventive maintenance of a production system has to be carried out during some limited time period outside regular production hours. Secondly, in some situations the repair shop capacity is flexible, e.g. because the repairmen work overtime if the workload is high or because it is possible to outsource service part repair if the workload is high. Using this flexible capacity, the repair shop has virtually infinite resources. However, working overtime may cause high overtime costs, while outsourcing is not always possible, in particular if the parts or modules are technically complicated and repair requires specific skills. Thirdly, the finite capacity could be taken into account by measuring the actual throughput time in the repair shop and plugging these values into the model as gross repair times (=net repair time plus waiting time). Although this seems to be a simple and straightforward solution, the throughput time measured is only valid under the current circumstances. Obviously the throughput times depend on service part demand, return procedures and repair shop capacity. Therefore this procedure is not suitable for what-if analyses, which is a serious drawback.

Another reason to include finite repair capacity in the model is the following. Under finite capacity, the item throughput times can be influenced using appropriate priority setting. For example, expensive items can be given high priority, so that repair shop throughput times are short and hence the stock levels required can remain low. This is only possible at the expense of lower priority for the other (cheaper) items, so that these items will face longer throughput times and hence require higher stock levels. If the item values are very different, such priority setting might be worthwhile to consider in order to improve the ratio between system availability and inventory investment. In order to be able to make such a trade-off, priority queuing models need to be incorporated in a METRIC-like approach. As a first step in this direction, we will focus on finite repair shop capacity under first-come-first-served rules in this paper.

Our choice to examine finite repair capacity is further motivated by Rustenburg et al. [5], who state that "one of the most criticized assumptions on METRIC-models and their extensions is the assumption of unlimited repair capacity". They put capacity restrictions on their agenda for research, but note that "the combination of capacitated multi-indenture and multi-echelon structures however will require a substantial research effort". We aim to contribute to this research effort with our paper.

We will use the VARI-METRIC method as starting point (cf. [6,4]). In the next section, we will first discuss the research results that are available in literature with focus on finite capacity. In Section 3, we will describe our model in detail. In Section 4, we will give a preliminary analysis of the impact of finite capacity. We will show that simply "plugging in" average throughput times in the VARI-METRIC model may lead to inaccurate results. Next, we will develop a method to optimise spare part stock levels under finite repair shop capacity, which is an extension of Slay's VARI-METRIC method (Section 5). In Section 6, we compare the system availability obtained from a numerical simulation for different input sets to the system availability estimated by presented approximation. Finally, we present our conclusions and possible directions of further research in the last section.

2. Literature overview

The service network model considered in this article describes the process in which operating units are sent to repair after failure, and after repair they return as good as new. Since the repair process in this model has several echelons of supply (local bases, central depots, etc.), and takes into account the product structure of failed units (assemblies, subassemblies), the model will be referred to as a *multi-echelon*, *multi-indenture* model.

In the literature, various approaches to solve such multi-echelon, multi-indenture models are described. As Guide and Srivastava [7] state, "the classical repairable problem is the military logistics problem of stocking repairable parts for aircrafts at bases which are capable of repairing some, but not all broken parts, and a central depot which serves all of the bases". The basic reference in this area is the METRIC method of Sherbrooke [3]. Models of the METRIC class have initially been applied to various military systems in the air force, the navy and the army [4]. This model class is generally recognized to be useful for commercial applications as well insofar repairable items are involved. Guide and Srivastava [7] mention "the leasing of office equipment (e.g. reproducing equipment and computers), heavy equipment (e.g. tractors, earth-moving equipment, industrial presses), and transportation equipment such as railroad, subway cars and buses". Practical applications using the METRIC method are described in various commercial settings, such as aircrafts [8], the Venezuelan metro-system [9] and electronic testing equipment [10]. For a general review of these kinds of models, we refer to Guide and Srivastava [7], Rustenburg et al. [5], and Kennedy et al. [11]. Because the classical application in the area are military systems, we will use cases derived from Rustenburg [12] and Sherbrooke [4] in this paper.

A well-known approach is the METRIC method developed by Sherbrooke [3]. This approach and its extensions (e.g. VARI-METRIC of [6]) employ simple assumptions that make it very easy to use in practice. One of these assumptions is commonly known as the ample service assumption. It means that the repair capacity is infinite, i.e. there is no queue of items waiting for a repair channel. This has the effect that the replenishment lead times can be considered as statistically independent, and the mean and variance of the number of items under repair service are equal. As discussed in the introduction, the infinite capacity assumption may not be applicable in practice. As a consequence, the independence of lead time does not apply anymore. This influences performance calculations (backorders, availability) as well the optimisation procedure.

Of course, the capacity effects have been recognised both in practice and in literature. Pyke [13] discusses the impact of finite capacity and repair priority setting using discrete event simulation. He finds that applying appropriate priority rules can have significant impact on the system performance if the utilisation is high. An analytical method is not developed, however. De Haas and Verrijdt [8] examine capacity effects and encounter heavily fluctuating work loads in the repair shops of the aircraft maintenance system they study, causing varying throughput times and frequent overtime and stress for the repair men. They suggest that the throughput times in the repair shop should be corrected for utilisation of repairmen. Although they consider various options for repair shop throughput times in their numerical illustration, they do not describe a method to quantify the relation between utilisation and throughput time. Hence the issue how to deal exactly with finite capacity remains unsolved.

In the literature, various ways to deal with finite capacity in service part networks have been discussed. One of these methods is to model the network as the closed queuing network (Jackson network, cf. [14,15]). This method provides very good estimations of the steady state probabilities in a closed network with fixed parameters, but the numerical algorithms involved make it difficult to find optimal stock levels for each location and each part type. Another approach is based on Markov processes; see [16–18]. A drawback of this approach is the fact that the number of states may become very large and that existing methods to reduce the model size to acceptable dimensions are rather rough.

A similar approach is developed by Avsar and Zijm [19]. They construct an excellent approximation for a two-echelon inventory model, where repair shops can be modelled as open Jackson queuing networks. However, their model considers only item-dedicated repair shops and is difficult to extend to multi-echelon model or model with different types of repair shops, as we consider.

Another possibility is to extend the VARI-METRIC method to deal with finite capacity by replacing the M/G/∞ queuing model for the repair shop by some finite capacity system, cf. [9,20,21]. They use their method to analyse the impact of finite capacity. They show that finite capacity has a serious impact on system performance for a single indenture, two-echelon system with only one central repair shop. They model the repair shop as a GI/G/k multi-class queuing system, where the part flow of one item type is modelled as one class in the queuing system. Although they discuss formulas for multi-server queues, their numerical results refer to single server queues only. In addition, they discuss an alternative method to plug in throughput times as observed in practice in the M/G/∞ model, so that waiting times are included. This approach is also used in the case study for the Caracas Metro subway system that Diaz and Fu [9] present. Then the impact of finite capacity is less, but still significant, and as we mentioned already this procedure is not suitable for what-if analyses.

Although the approach by Diaz and Fu [9] is simple and attractive, they restrict themselves to a very simple situation, namely a single repair shop consisting of one or more single server queues for a single-indenture, two-echelon system. It can be expected that the model complexity increases if the service part supply system contains multiple repair shops at various locations in the network and for various levels in a multi-indenture system, because then the various queuing models interact which may cause a serious deterioration of numerical accuracy. Also, the model performance for multi-server repair shops has not been analysed. In this article, we generalise the model of Diaz and Fu [9] to the multi-echelon, multi-indenture systems that may have repair shops at multiple locations in the network (local/central). To this end, we use the approach by van Harten and Sleptchenko [22] to compute performance characteristics of multi-class, multi-server queuing systems. Our aim is to get insight into the impact of capacity restrictions on the system performance in terms of the expected number of backorders, and system availability. Also, we will examine how the service part allocation in the network is affected by the finite capacity. In other words, to which extent will the use of the infinite capacity assumption lead to sub-optimal decisions (lower availability for the same budget).

3. The model

In this section, we first describe Sherbrooke's model for a general multi-item, multi-indenture, multi-echelon inventory system (Section 3.1). We introduce our extension of this model to repair shops with finite capacity in Section 3.2. The model assumptions and basic notation are presented in Section 3.3. As we will proceed from the VARI-METRIC method (cf. [6]), we give an outline of this method in Section 3.4.

3.1. Multi-echelon, multi-indenture systems

We show an example of both a multi-echelon supply network and a multi-indenture product tree in Fig. 1 (taken from [4]). The installed base consists of the pumps aboard the submarines, each consisting of several modules and parts (figure on the right). The numbers at the connectors represent the probabilities that the failure of a system/module is caused by a certain item. The echelon structure consists of a depot, 4 supply ships and 16 submarines (figure on the left). If an item cannot be repaired at the most upstream location, it can be sent to an external supplier for repair or replacement. The internal structure of the external supplier is not explicitly considered, but just modelled as a "black box" with corresponding throughput time characteristics.

3.2. Repair shops

We define a repair shop as a part of a location that has its own repair facilities. We consider two variants of repair facilities:

- An item-dedicated repair facility, that is able to repair only one kind of items. Then every

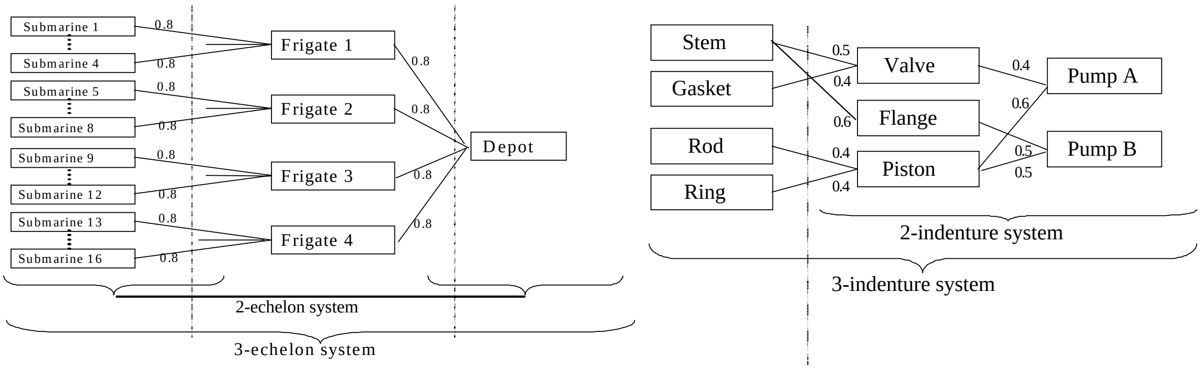

Fig. 1. An example of a supply system and an item hierarchy.

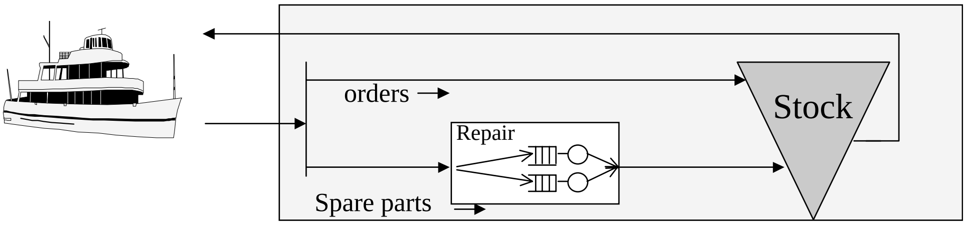

Fig. 2. Single-site, single-indenture system (one location consisting of a stockpoint and a repairshop).

location has as many repair facilities as the number of items.

- A cluster-dedicated repair facility, when multiple kinds of items can share this facility. Here we define a *cluster* as a set of items, which can be repaired at the same facility using the same shared repair resources (personnel, equipment). There is no limitation in the composition of the clusters.

These two variants imply different queuing models: *single-class* multi-server queue, for an item-dedicated repair facility, and a *multi-class* multi-server queuing model for cluster-dedicated cases. The repair shops for both variants of repair facilities have similar structures. Fig. 2 presents an example of a single-site, single-indenture system.

This simple system consists of only one location having repair facilities and only single level indenture items circulate in this system. The repair block in Fig. 2 may consist of one or more dedicated repair facilities. Items are processed by a repair shop as follows: First, a failed item arrives at the repair shop and an order for a new item is issued. If a new item is available on stock, it is dispatched to replace the failed item. At the same time, the failed item enters the repair facility and after repair it is added to stock. The number of items in this sequence of processes between arrival of the failed item at a repair shop and repair completion is called the *pipeline*.

In a *multi-echelon* structure, the failed item can follow two different routes through the system: either it enters the local repair facility, or it is forwarded to the next echelon upstream to be repaired there (e.g. the downstream echelon consists of supply ships, the upstream echelon consists of a depot in the harbour, see Fig. 1). Usually items are sent to the higher echelon if local repair is technically impossible, i.e. if the local repair shop does not have appropriate equipment or skills. Sherbrooke [4] assumes that the decision whether to repair locally or not is based on such technical considerations only, and not by, e.g.

current repair shop workload. This is modelled by a fixed probability that the item can be repaired locally, independent of the system state. We will use the same assumption.

A *multi-indenture* structure means that every item (assembly) may consist of other items (subassemblies), see Fig. 1. If an assembly fails, we assume (following [4]) that either the failure is caused by the failure of exactly one of its subassemblies (so the replacement of this subassembly is sufficient to repair the assembly), or there is no specific subassembly causing the assembly failure (hence the assembly as a whole has to be repaired). In the case of a subassembly failure, the repair procedure is as follows: the item is disassembled and the failed subassembly is sent to the subassembly repair shop, where the procedure is similar: First, a new item is ordered from stock, and then the failed item enters repair.

3.3. Assumptions and notation

In this paper, we use the following assumptions:

1. Demands occur according to stationary Poisson processes, independent of the number of items under repair (i.e. the impact of the finite installed base is neglected).
2. The failure of an assembly is caused by at most one subassembly failure.
3. Each stockpoint uses an (S−1, S) inventory policy for each item, i.e. the stock level equals S and each demand immediately generates an order for a replacement item; as a consequence, there is no batching.
4. Each stockpoint has exactly one supplier; there is no lateral supply from other stockpoints.
5. Request for replacement items are handled by a stockpoint according to the first come, first serve (FCFS) rule (i.e. no allocation priorities).
6. After repair, the items are as good as new.
7. Backorders for different items are equally important.
8. The item repair times are independent, identically distributed random variables.
9. The repair shops handle their jobs according to the FCFS discipline (i.e. no repair priorities).
10. The probability that an item is repaired in a particular repair shop is solely determined by technical considerations and not by the system state (e.g. repair shop utilisation); this is by a fixed repair probability for each combination of item and location.

Assumptions 1–8 are generally used in the literature on repairable spare part management (cf. [4,12,5]). Some specific models have been developed to examine the impact of releasing some of these assumptions. For example, Verrijdt [2] examines lateral supply (releasing assumption 4) and concludes that this has only significant impact for low fill rates at downstream locations (<70%). This justifies assumption 4. Pyke [13] examines dispatch policies using a simulation model (releasing assumption 5) and concludes that his distribution rule has little effect for most cases. This justifies using assumption 5. For a more extensive motivation and critical review of assumptions 1–7, we refer to Rustenburg [12].

Modelling finite capacity repair shops imply that we have to make assumptions on the structure and processes within the repair shop. Therefore we need some additional assumptions (8–10). Regarding assumption 8, it is in fact different from the commonly used assumption of independent repair *lead* times. If the repair shop capacity is infinite, the throughput times of consecutive jobs are generally assumed to be independent because of the absence of a practical alternative [12]. Finite capacity repair shops automatically deal with *dependent* repair lead times, because the waiting times of consecutive jobs are naturally correlated. In this sense, our assumption is less restrictive than the one commonly used in the literature. If the times to carry out repair jobs (excluding waiting times) would be correlated as well, we would have to deal with: (1) an intractable queuing model for which no useful results are available in the literature, (2) a practical problem because these correlations are hard to estimate from field data.

The queuing discipline (FCFS) is an important characteristic of a finite capacity repair shop. Pyke

[13] examined the impact of repair priorities using his simulation model and found that priority rules are clearly valuable. This requires (multi-class, multi-server) priority queuing models which complicates the mathematical analysis considerably. Therefore we choose to use the FCFS discipline in order to get more insight into the impact of finite repair capacity. Knowing that appropriate priority rules can enhance the system performance indeed, a future extension of our model to priority repair is high on our research agenda.

In fact, assumption 10 on repair job routing is generally used in the literature on spare part networks with infinite capacity repair shops, similar to assumptions 1–7 [4,12]. This is not surprising, because the infinite capacity assumption rules out the option to route repair jobs through the network based on repair shop utilisation. Besides, technical considerations can be dominant indeed, because expensive repair equipment and advanced technical skills tend to be concentrated in a central repair shop. In such a situation, a local repair facility is only able to carry out relatively simple jobs because equipment is not available. In situations where such restrictions do not occur, dynamic routing of repair jobs based on time-dependent repair shop utilisation is an option. Because of the additional transport costs involved if a failed item has to be moved to another location, we expect that releasing assumption 9 will have more perspective for efficiency gain than assumption 10.

In the remainder of this paper, we will use the following notation:

Geographical structure

m = location index, $m = 1, \ldots, M$, where M denotes the total number of locations in the system

n = echelon index, $n = 1, \ldots, N$, where N denotes the total number of echelons in the system; here $n = 1$ ($n = N$) denotes the most upstream (downstream) level

ECH(n) = set of locations belonging to the echelon n

CUS(m) = set of customers (supported locations) of location m

SUP(m) = the location that supplies items to location m (i.e. the direct supplier of location m)

Product structure

j = item index, $j = 1, \ldots, J$, where J denotes the total number of items in the system

i = indenture index, $i = 0, \ldots, I$; here $i = 0$ denotes the system level and $i = I$ denotes the lowest subassembly level, i.e. the subassemblies that cannot be decomposed further into smaller units

IDN(i) = set of items belonging to indenture i

SA(j) = set of subassemblies of item j

AS(j) = set of assemblies, which have item j as a subassembly

Z_j = the multiplicity of item j (i.e. the number of class j items in a single system)

Repair shop characteristics

λ_{mj} = demand rate for item j at location m

S_{mj} = repair time, a random variable with mean $E[S_{mj}]$ and coefficient of variation C_{Smj}, where the coefficient of variation is defined as the ratio between the standard deviation and the mean

Q_{mj} = number of type j items in the repair queue at the location m

R_{mj} = number of type j under repair (in queue and in service) at location m

Routing characteristics

r_{mj} = probability that item j can be repaired at location m; the item is dispatched to the next higher level in the multi-echelon structure with probability $(1 - r_{mj})$

q_{mjk} = probability that a failure of item j at location m is caused by a failure of subassembly k ($k \in \mathrm{SA}(j)$)

Other notation

B_m = the size of the installed base at location m

s_{mj} = stock level of item j at location m; $\bar{s}$ denotes the matrix of all stock levels in the system

P_{mj} = number of type j items in the pipeline at the location m, i.e. all items j under repair at location m, all items j on order at the supplier of location m and all items j waiting at location m for subassembly replacement

O_{mj} = order-and-ship time of item j to location m from its supplier, a random variable with mean $E[O_{mj}]$ and coefficient of variation $C_{O_{mj}}$

$\mathrm{BO}_{mj}(\bar{s})$ = number of backorders for item j at location m as function of the stock levels $\bar{s}$

$\mathrm{PBO}_{mj}(\bar{s})$ = backorder probability for item j at location m as function of the stock levels $\bar{s}$

c_j = price of item j

The decision variables in our model are the stock levels s_{mj}.

3.3.1. Outline of VARI-METRIC

There are different ways to measure availability of supply systems for repairable items. It is common to relate the average system availability to the backorders [4]:

$$A \approx \frac{1}{|\mathrm{ECH}(N)|} \times \sum_{m \in \mathrm{ECH}(N)} E\left[\prod_{j \in \mathrm{IND}(1)} \{1 - \mathrm{BO}_{mj}(\bar{s})/(B_m Z_j)\}^{Z_j}\right], \tag{1}$$

where $|\mathrm{ECH}(N)|$ denotes the number of elements in the set $\mathrm{ECH}(N)$. Hence the average system availability depends on the backorders of all highest indenture items ($j \in \mathrm{IND}(1)$) at all downstream locations ($m \in \mathrm{ECH}(N)$). Sherbrooke [4] shows that maximising this availability function is approximately equivalent to minimising the sum of the expected backorders if the repair shops have *infinite* capacity. This leads to the following goal function:

$$\min_{\bar{s}} \sum_{m \in \mathrm{ECH}(N)} \sum_{j \in \mathrm{IND}(1)} E[\mathrm{BO}_{mj}(\bar{s})]. \tag{2}$$

Another availability function was introduced by Rustenburg [12] in a NAVY case. In this application, the downstream locations are frigates and the failure of any unit implies that the frigate is not available. In fact, the size of the installed base at each downstream location is exactly one ($B_m = 1$). For this case the availability function can be defined as

$$A = \frac{1}{|\mathrm{ECH}(N)|} \times \sum_{m \in \mathrm{ECH}(N)} \prod_{j \in \mathrm{IND}(1)} \{1 - \mathrm{BO}_{mj}(\bar{s})\}. \tag{3}$$

Rustenburg [12] shows that maximising this availability function is approximately equivalent to minimising the sum of the backorder probabilities. So an alternative goal function is

$$\min_{\bar{s}} \sum_{m \in \mathrm{ECH}(N)} \sum_{j \in \mathrm{IND}(1)} \mathrm{PBO}_{mj}(\bar{s})]. \tag{4}$$

The optimisation procedure is similar for both goal functions. Both for the traditional infinite capacity repair shops and for our finite capacity models it is possible to calculate both expected backorders and backorder probabilities. Therefore, we can use our model for various availability functions. We will return to the methods of estimating the system availability in the case of *finite* capacity repair shops in Section 5.2.

Next, we will discuss how to find the distribution of backorders in the system. Consider a subsystem with one repair facility and one stock at location m in a multi-echelon multi-indenture system. The number of backorders of item j in such subsystem is the non-negative difference between number of items in pipeline P_{mj} and the number of items in stock s_{mj}:

$$\mathrm{BO}_{mj} = \max\{P_{mj} - s_{mj}, 0\}. \tag{5}$$

Recall that the number of items in the pipeline refers to items in local repair plus backordered items in external repair, where external repair means repair at a higher echelon or repair of subassemblies. Hence the pipeline distribution depends on the backorders upstream in the echelon structure and downwards in the indenture structure. The (VARI-)METRIC method analyses the relevant backorder distributions by subsequently deriving pipeline and backorder distributions of various items and various locations, starting from the lowest indenture items at the most upstream location in the echelon structure. To facilitate this, the arrival rates of all items (assemblies and subassemblies) at all repair shops in the network are calculated.

Below, we give an outline of the basic mathematics of VARI-METRIC.

3.4. Calculating the item arrival rates

The arrival rate $_{mj}$ of item j at location m can be derived from two parts:

- the arrival rates of this item at downstream locations $l(\lambda_l)$ multiplied by the probability that these items cannot be repaired at these downstream locations $(1 - r_{lj})$;
- the arrival rates of higher indenture items (i.e. assembles at k have item j as subassembly) at location $m(\lambda_{mk})$ multiplied by the probability that item k can be repaired at location $m(r_{mk})$ and multiplied by the probability that item j is the cause of the failure (q_{mkj}).

Hence we arrive at the expression

$$\lambda_{mj} = \sum_{l \in \mathrm{CUS}(j)} \lambda_{lj}(1 - r_{lj}) + \sum_{k \in \mathrm{SA}(j)} \lambda_{mk} r_{mk} q_{mkj}. \qquad (6)$$

All arrival rates can be computed recursively, starting from the failure rates of systems (the highest indenture items) in the operational field (the most downstream locations in the echelon structure).

3.5. Calculating pipeline and backorder characteristics

The dependency of backorders on the pipeline is given by (5). Next to this dependency, we have a relation between the pipeline and the backorders for other combinations of locations and items. We first state this dependency for the mean number of items in the pipeline, next we will give the expression for the variance. Referring to the definition of pipeline, we see that the number of items of type j in the pipeline at location m consist of:

- *All items j under repair at location m.* If the repair shop has infinite capacity, the mean number of items under repair is given $\lambda_{mj} r_{mj} E[S_{mj}]$.
- *All items j waiting at location m for subassembly replacement*. The reasoning to derive an expression for this number is as follows: the total number of items waiting for replacement of subassembly $k \in \mathrm{SA}(j)$ equals the number of backorders BO_{mk}. Only a fraction h_{mjk} of the backorders for item k at location m is due to a request from item j. A reasonable approximation for this fraction is the effective demand rate for subassembly k arising from item j as a fraction of the total demand rate for item k at location m: $h_{mjk} = r_{jm}\lambda_{mj}q_{mjk}/\lambda_{mk}$. So for this part of the pipeline, we arrive at the mean value $\sum_{k \in \mathrm{SA}(j)} h_{mjk} E[\mathrm{BO}_{mk}]$.
- *All items j on order at the supplier of location m.* These items include the items being transported to location j from its supplier $(\lambda_{mj}(1 - r_{mj})E[O_{mj}])$ and the items waiting at the supplier $n = \mathrm{SUP}(m)$ for replacement. The items waiting for replacement can be derived from the number of backorders for item j at the supplier n, BO_{nj}. Only a fraction f_{mj} of these backorders are destined for location m. This fraction can be calculated as the ratio between the demand for item j by location m and the total demand for item j at the supplier n: $f_{mj} = (1 - r_{jm})\lambda_{mj}/\lambda_{\mathrm{SUP}(m),j}$. So for this part of the pipeline, we arrive at the mean value $\lambda_{mj}(1 - r_{mj})E[O_{mj}] + f_{mj}E[\mathrm{BO}_{\mathrm{SUP}(m),j}]$.

Putting it all together, we find the following expression:

$$E[P_{mj}] = \lambda_{mj} r_{mj} E[S_{mj}] + \sum_{k \in \mathrm{SA}(j)} h_{mjk} E[\mathrm{BO}_{mk}] + \lambda_{mj}(1 - r_{mj})E[O_{mj}] + f_{mj}E[\mathrm{BO}_{\mathrm{SUP}(m),j}]. \qquad (7)$$

Eq. (5) shows that the backorder characteristics can be calculated from the pipeline characteristics, whereas Eq. (7) shows that the pipeline characteristics can be calculated from the backorder characteristics. This provides a way to obtain all backorder characteristics recursively. As a simple approximation to obtain the backorder characteristics from (5), we only use the first two moments of the numbers of items in the pipeline and we fit a discrete distribution on the first two moments. The most general way to achieve this is given by Adan et al. [23]. Based on this approximate distribution for the pipeline, we can find the first two moments of the backorders. Next we can use the values for the mean backorders to calculate the mean pipeline for higher indenture items and for more downstream locations from (7). Similar to the expression for the mean backorders, an expression for the variance of the backorders can be derived as well. Then we arrive at (cf. [4,12]):

$$\begin{aligned}\mathrm{Var}[P_{mj}] = {} & \lambda_{mj} r_{mj} E[S_{mj}] \\ & + \sum_{k \in \mathrm{SA}(j)} \{h_{mjk}(1-h_{mjk})E[\mathrm{BO}_{mk}] \\ & + h_{mjk}^2 \mathrm{Var}[\mathrm{BO}_{mk}]\} + \lambda_{mj}(1-r_{mj})E[O_{mj}] \\ & + f_{mj}(1-f_{mj})E[\mathrm{BO}_{\mathrm{SUP}(m),j}] \\ & + f_{mj}^2 \mathrm{Var}[\mathrm{BO}_{\mathrm{SUP}(m),j}]. \qquad (8)\end{aligned}$$

3.6. *The VARI-METRIC algorithm*

Based on the equations for the item arrival rates and the backorder characteristics, the VARI-METRIC algorithm uses a greedy algorithm to optimise item stock levels within a budget constraint. The idea is very simple: The system availability is related to the backorders of all highest indenture items ($i \in \mathrm{IDN}(1)$) at almost all downstream locations ($m \in \mathrm{ECH}(N)$). Hence a greedy algorithm adds an item i^* to stock at location m^* in each step ($\bar{s} + e_{i^* m^*}$, where e_{im} is a matrix with all elements equal to zero, except for element i, m which is equal to 1), such that the reduction in the sum of the expected backorders per invested dollar is maximised. So for goal function (2), the rule is to select (i^*, m^*) such, that the expression

$$\Delta_{i^* m^*} = \left\{ \sum_{j \in \mathrm{IDN}(1)} \sum_{m \in \mathrm{ECH}(N)} E[\mathrm{BO}_{mj}(\bar{s})] - \sum_{j \in \mathrm{IDN}(1)} \sum_{m \in \mathrm{ECH}(N)} E[\mathrm{BO}_{mj}(\bar{s} + e_{i^* n^*})] \right\} \Big/ c_{i^*} \qquad (9)$$

is maximised. For goal function (4), we simply substitute $E[\mathrm{BO}_{mj}(\bar{s})]$ by $\mathrm{PBO}_{mj}(\bar{s})$. Sherbrooke [4] shows that this greedy method yields an optimal solution for single-site, multi-product, single-indenture models. The algorithm is not optimal for general multi-echelon, multi-indenture models. However, Rustenburg [12] shows that this greedy algorithm yields good results, provided that some nonnegative starting values are chosen for the stock levels. We recall that the expressions as presented in this section are only valid for infinite capacity repair shops. For that model, the variance of the number of items under repair equals the mean value, as is well known from the analysis of the $M/G/\infty$ queue. For the modifications of the VARI-METRIC algorithm to the finite capacity model, we refer to Section 5.

4. A preliminary analysis of the impact of finite capacity

As mentioned in the introduction, a simple approach to deal with finite repair capacity is to measure throughput times and plug these values as repair times in the (VARI-)METRIC model. However, this may lead to inaccurate intermediate results, as we will show in this section. The issue is that the aforementioned approach does not guarantee that the *variance* of the number of items under repair is correctly estimated.

To this end, we compare two models. In the first model, we have a finite capacity repair shop where the part throughput time consists of waiting time and repair time. A single repair shop handles all items. Assuming for this simple example that the repair times are the same for all items, we can model the repair shop as an $M/M/k$ *queue*. In the second model, we assume an infinite capacity

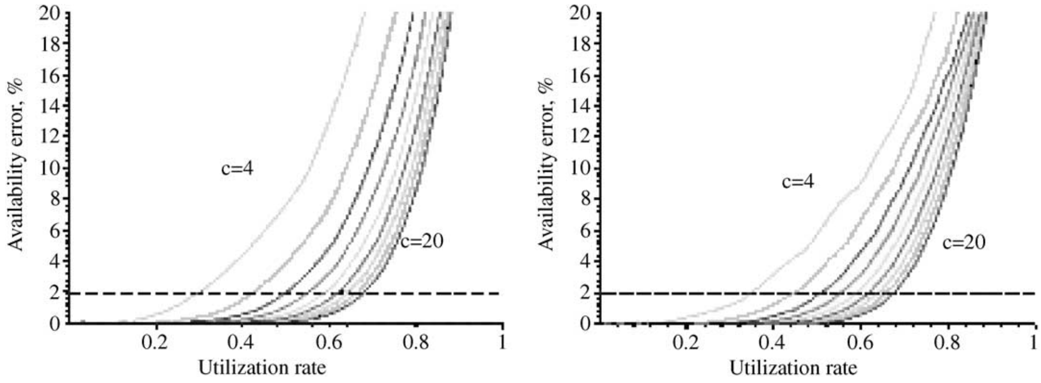

Fig. 3. The availability errors for different numbers of servers and different utilisation rates (one- and two-echelon systems).

repair shop, modelled as an M/G/∞ queue where the service time equals the sum of the waiting time and repair time resulting from the finite capacity system as found from the first model. Then, the number of parts under repair, in the second model, is Poisson distributed, so that the variance of the number of parts in the repair shop equals the mean. Obviously, the latter need not be true in the first model. Hence the error made by using a simple M/G/∞ queuing system is caused by misspecification of the *variance* of the number of items in the queuing system.

To compare these models for repair shops, we analyse simple single models for both single-echelon and two-echelon models. The two-echelon model consists of one central depot and one local base. We assume that *all* repair shops have finite capacity and that the utilisation rates and number of servers are equal for all repair shops. For these models, we first find the mean number of items in service based on utilisation rate and number of servers. Then, we optimise the stock levels according to the VARI-METRIC method, where the variance of number of items in repair is equal to the mean of this number for different availability levels, from 0% to 100%. Thereby we estimate the system availability using the infinite capacity assumption. Also, we estimate the system availability based on a *finite* capacity repair shop for each combination of stock levels, as follows. We can easily find the variance of the number of items in the repair shop, proceeding from an M/M/k queue. This variance, which is generally different from the mean, is used to improve the estimate of the system availability using the VARI-METRIC method.

Using the results for two simple systems (one echelon and two echelon), we can find the maximum availability error (absolute difference in availability between M/G/k and M/G/∞ repair shop) corresponding to every combination of number of servers and utilisation rate. Then, we can make plots (Fig. 3), which show us for every number of servers and for every utilisation rate the maximum error in the system availability if we ignore finite repair capacity.

These plots show us the critical utilisation rates in simple systems. For example, if one tolerates an error of 2% points in availability for a two-echelon system with $c = 4$ servers, the maximum acceptable utilisation rate to use VARI-METRIC is approximately 0.45 or less (all values for which left function on every plot is below the dashed line).

5. Stock allocation under finite capacity

It has been shown in the previous section that ignoring finite capacity can have a significant influence on the estimate of the availability. Therefore we will develop a variant of the VARI-METRIC method, taking into account finite capacity of repair shops.

Sherbrooke [4] shows that the expected number of backorders depends on the mean and variance of the number of items in the pipeline. To calculate these pipeline characteristics, we need the mean and variance of the number of items in the repair shop. If the repair shop is modelled as M/G/∞ queue, the number of items in the repair shop is Poisson distributed. However, such a simple expression is not available for single-class, multi-server queues in case of a general service time distribution (if repair shops are item-dedicated) or multi-class multi-server queues (if repair shops are cluster-dedicated). There are only few useful results from literature, because most queuing theory focuses on the mean and (sometimes) variance of the number of items *in queue* rather than *in the system* (in queue plus in service). It is hard to derive the variance of the total number of items in the system from these results, because (a) the variance of the number of items in repair is not known and (b) the numbers of items in queue and in repair are *not* independent. For the multi-class, multi-server M/M/k queue, we can use the results from Sleptchenko and van Harten (2000). This will be discussed in Section 5.1. Obviously, we can use these results for the *single*-class, multi-server M/M/k queue as well. Then, we discuss the impact of finite capacity on the optimisation procedure in Section 5.2.

5.1. Multi-class M/M/k queue

The repair shops at the lowest echelon are usually run by a small crew of multi-skilled specialists that are able to handle a certain set of repair jobs. This gives rise to multi-class multi-server models, where several classes of items with each their own arrival and service processes share the same queue and the same servers. A complication of this model is that the number of items in the system are *correlated* amongst classes. If the number of class j items is high at a particular point in time, it is likely that the servers had a high utilisation recently, and so the number of class $j' \neq j$ items in the system can expected to be high as well. The consequences of this phenomenon on the optimisation procedure will be discussed in Section 5.2.

Multi-class, multi-server queuing models have not been investigated as extensively as single-class queuing models. Still some results are worth mentioning. Diaz and Fu [9] present an approximation for the GI/G/*1* multi-class queue, but they do not discuss the multi-server model. The most advanced theoretical results are obtained by de Smit [24] for the multi-class GI/M/k model (or, equivalently, the single class GI/H/k model where H denotes the hyperexponential distribution). However, though his approach leads to nice structural properties of the waiting time distribution, it cannot provide the sort of information on performance characteristics concerning backorders that we need. Moreover, the approach does not look promising from a numerical point of view as well and, as far as we know, it was not tested numerically.

A more useful approach is given by Van Harten and Sleptchenko [22], who develop both an exact solution and an approximate variant for the multi-class M/M/k queue. In this paper, we will use their approximation, which is quite easy from a computational point of view and gives accurate results for high utilisation rates. As can be seen from Section 4, this is the most interesting situation. For low utilisation rates, we might as well use the M/G/∞ queue. The main feature of the multi-class queue is that all items of one cluster C are sharing the same repair facility, which has k_C servers and a utilisation rate ρ_C. Similar to Van Harten and Sleptchenko [22], we define the relative arrival rate a_{jm} of item j at location m as

$$a_{mj} = \frac{\lambda_{mj}}{\Lambda_{\mathrm{C}}} \text{ where } \Lambda_{\mathrm{C}} = \sum_{(j') \in C} \lambda_{mj'}.$$

The relative deviation of the service rate of item j at location m from its cluster average is denoted by the variable δ_{mj}, i.e.

$$\delta_{mj} = \frac{E[S_{\mathrm{C}}]}{E[S_{mj}]} - 1,$$

$$\text{where } E[S_{\mathrm{C}}] = \frac{1}{\Lambda_{\mathrm{C}}} \sum_{(m',j') \in C} \lambda_{m'j'} E[S_{m'j'}].$$

As the analysis focuses on a repair shop for a single cluster of items at a single location, we simplify the notation by omitting the cluster index C and the location index m. So we will work with an abstract queue with k servers and utilisation rate ρ. The queue is shared by $|C|$ (number of items in cluster C) items having arrival and service time characteristics $a(j)$ and $\delta(j)$, respectively ($j = 1, \ldots |C|$).

To analyse the system as a Markov chain, the system state is described by two vectors w and v where the ith element w_i of the vector w gives the class of the *i*th item in queue and the *i*th element v_i of the vector v gives the class of the item in service at server i $(i = 1, 2, \ldots, k)$. This abundant state representation facilitates the solution of the equilibrium equations. The approximate solution is based on constructing a product form solution of the equilibrium equations. It first gives an approximation for the system state probabilities, and next these probabilities are used to approximate the necessary performance measures such as mean and variance of number of items in the system for each class j ($E[R_j]$ and $\mathrm{Var}[R_j]$). We can even approximate correlations between R_i and R_j $(i \neq j)$ from the state probabilities.

The product form of the probability on system state (w, v) is

$$P(w,v) = \left(\sum_{j=1}^{|C|} w_j\right)! \prod_{j=1}^{|C|} \frac{a(j)^{w_j}}{w_j!} P_R(v),$$

where $P_R(v)$ denotes the probability on server state v given that the total number of items in the system equals R.

The components $P_R(v)$ of the product form solution can be calculated directly when the queue is empty:

$$P_R(v) = p_0 \frac{(k\rho)^R}{R!} \binom{k}{R}^{-1} \times \prod_{v(j) \neq 0} \frac{a(v(j))}{\{1 + \delta(v(j))\}}, \quad R \leqslant k. \tag{10}$$

The probabilities of the other states can be calculated using the eigenvalues z_i $(i = 1 \ldots |C|)$ of the linear system of equilibrium equations describing the queuing process (cf. [22]). Although this procedure gives exact probabilities of the system states, the procedure of finding eigenvalues may require an excessive amount of computation time if the number of classes and servers is high. As an approximation, we observe that some eigenvalues can be calculated as

$$z_i(\sigma_i) = (1 + \rho - \sigma_i)/\rho,$$

where σ_i is the *i*th solution of the equation

$$1 - \rho - \sigma = \rho \sum_{j=1}^{N} a(j) \frac{1 + \delta(j)}{(\sigma + \delta(j))}.$$

These equations give exactly $|C|$ eigenvalues > 1 and one eigenvalue $= 1$ (cf. [22]). Compared to the standard procedure for finding eigenvalues, these equations provide a considerable easier method from a computational point of view. Therefore, we can construct an approximation for the probabilities $P_R(v)$ based on the $|C|$ eigenvalues > 1, z_i and their corresponding eigenvectors:

$$P_R(v) \approx \sum_{i=1}^{|C|} \gamma_i z_i^{-R+k} D(\sigma_i)^{-k} \times \prod_{j=1}^{|C|} \frac{a(v(j))}{\{1 + \delta(v(j))/\sigma_i\}}, \quad R > k, \tag{11}$$

where $D(_i)$ is defined by

$$D(\sigma_i) = \sum_{i=1}^{|C|} \frac{a(l)}{\{1 + \delta(l)/\sigma_i\}}.$$

The coefficients p_0 in expression (10) and γ_i $(i = 1 \ldots |C|)$ in expression (11) should be chosen such that the expected number of class j items in service is equal to $\lambda(j)/\mu(j) = \rho(j)$ and the sum of all probabilities is equal to 1. Thus, we obtain a set of $|C|$ equations:

$$\sum_{i=1}^{|C|} \left\{ D(\sigma_i)^{-1} \frac{a(j)}{\{1 + \delta(j)/\sigma_i\}} \right\} \varphi_i + p_0 b(j) \sum_{n=1}^{k-1} \frac{n}{k} \frac{(k\rho)^n}{n!} = \rho(j) \tag{12}$$

plus the sum of all probabilities:

$$\sum_{i=1}^{|C|} \varphi_i + p_0 \sum_{n=0}^{k-1} \frac{(k\rho)^n}{n!} = 1, \tag{13}$$

where the variables φ_i are defined as

$$\varphi_i \underset{\text{def}}{=} \gamma_i (1 - z_i^{-1})^{-1}.$$

Eqs. (12) and (13) give us a system of $|C| + 1$ linear equations in $|C|$ variables φ_i and one variable p_0. This system can be reduced to a $|C||C|$ linear system using the equation (see [22]):

$$p_0 = \left\{ \sum_{n=0}^{k-1} \frac{(k\rho)^n}{n!} \left(1 - \frac{n}{k}\right) \right\}^{-1} (1 - \rho).$$

Together, Eqs. (11)–(13) define an approximation for the system state probabilities. From these probabilities, we can derive approximations for the mean and the variance of the number of items in the queue for each class.

To calculate the mean value we have to find the sum in the following form:

$$E[R_i] = \sum_{R=1}^{\infty} \sum_{r=1}^{R} r \sum_{q,s} P_m(q, s|s_i = r)$$

which can be divided into two terms ($R<k$ and $R\geqslant k$). The first term is a finite sum of probability states that can be calculated directly from formula (3). The second term is in fact an infinite series of state probabilities, to be approximated by (4). After some algebra, we find the following approximation for the mean number of class i items in the system:

$$\begin{aligned} E[R_i] \approx\, & p_0 \sum_{R=1}^{k-1} \frac{(k\rho)^R}{R!} \binom{k}{R}^{-1} \\ & \times \sum_{r=1}^{R} r \sum_{s:s(j)=r} \prod_{s(j)\neq 0} \frac{a(s(j))}{\{1+\delta(s(j))\}} \\ & + \sum_{s:s(j)=R} s_i \sum_{i=1}^{N} \gamma_i (z_i - 1)^{-1} z_i D(\sigma_i)^{-k} \\ & \times \prod_{j=1}^{N} \frac{a(s(j))}{\{1+\delta(s(j))/\sigma_i\}}. \end{aligned}$$

Analogously we can approximate the second moment of the squared number of items in the system:

$$\begin{aligned} E[R_i^2] \approx\, & p_0 \sum_{R=1}^{k-1} \frac{(k\rho)^R}{R!} \binom{k}{R}^{-1} \\ & \times \sum_{r=1}^{m} r^2 \sum_{s,s.ts(j)=r} \prod_{s(j)\neq 0} \frac{a(s(j))}{\{1+\delta(s(j))\}} \\ & + 2 \sum_{s,s.t.} s_i \sum_{i=1}^{N} \gamma_i (z_i - 1)^{-2} z_i D(\sigma_i)^{-k} \\ & \times \prod_{j=1}^{N} \frac{a(s(j))}{\{1+\delta(s(j))/\sigma_i\}}. \end{aligned}$$

From the two equations above, we can find the variance using $\mathrm{Var}[R_i] = E[R_i^2] - E[R_i]^2$.

Van Harten and Sleptchenko [22] tested these approximations extensively and compared the approximate results to the exact results that they derived. They show that the approximation error remains within reasonable bounds ($<10\%$). Besides, they observe that the approximation error decreases with the repair shop utilisation. The latter is relevant, because our method is especially meant for instances with high utilisation.

5.1.1. Remark

An interesting generalisation is to combine these results for the multi-class M/M/k queue with the approximate equations that Whitt [25] developed from the relation between single-class M/M/k and G/G/k systems. This approximation does not make a deep analysis of arrival and repair process and the only characteristics of this processes used in this approximation are squared coefficients of service and interarrival time ($C^2_{A_{mi}}$ and $C^2_{S_{mi}}$) of item i at location m. In this way, we arrive at the following new approximations for the multi-class G/G/k queue:

$$\begin{aligned} E[R_{mi}]_{\mathrm{GI/G}/k} & \approx \left(\frac{C^2_{A_{mi}} + C^2_{S_{mi}}}{2} \right) E[Q_{mi}]_{\mathrm{M/M}/k} \\ & \quad + k_{mj}\rho_{mj}, \\ \mathrm{Var}[R_{mi}]_{\mathrm{GI/G}/k} & = E[R^2_{mi}]_{\mathrm{GI/G}/k} \\ & \quad - (E[R_{mi}]_{\mathrm{GI/G}/k})^2 \end{aligned} \tag{14}$$

and

$$\begin{aligned} E[R^2_{mi}]_{\mathrm{GI/G/k}} & \approx \left(E[R^2_{mi}]_{\mathrm{M/M}/k} / E[R_{mi}]^2_{\mathrm{M/M}/k} \right) \\ & \quad \times E[R_{mi}]_{\mathrm{GI/G}/k^2}, \end{aligned}$$

where the characteristics of multi-class M/M/k queue as described in this section are substituted. This approximation still has to be tested, however.

5.2. Optimisation procedure

We will base our optimisation procedure on the VARI-METRIC method as introduced by Sherbrooke [4]. Our motivation for using this approach is as follows. We deal with a *nonlinear* optimisation problem with many integer decision variables. This combination yields a hard optimisation problem. This is especially true for the finite capacity model that we consider, because we have to solve a multi-class, multi-server queuing model many times. Exact optimisation algorithms are only available for special cases, for example the single-item case [20]. Therefore we have to consider approximation schemes. Because it has been shown that the VARI-METRIC method provides good results if some appropriate starting values are chosen (cf. [12]), we decided to use this procedure. We only made a few modifications because of the finite capacity repair shops. Let us now discuss the concept of this procedure in more detail.

We proceed from the definition of the system availability as introduced by Sherbrooke [4], see Eq. (1). If the term $\mathrm{BO}_{mj}(\bar{s})/B_m Z_j$ is small, we can this equation by using the first order approximation $(1-\varepsilon)^k \approx 1 - k\varepsilon$ if ε is small. Then we find

$$A_m \approx E\left[1 - \sum_{j \in \mathrm{IND}(1)} \frac{\mathrm{BO}_{mj}(\bar{s})}{B_m}\right] = 1 - \frac{1}{B_m} \sum_{j \in \mathrm{IND}(1)} E[\mathrm{BO}_{mj}(\bar{s})] \tag{15}$$

and the average availability A over all systems at all downstream locations equals

$$A = 1 - \frac{1}{|\mathrm{ECH}(N)|} \sum_{m \in \mathrm{ECH}(N)} A_m \approx 1 - \frac{1}{|\mathrm{ECH}(N)|} \sum_{m \in \mathrm{ECH}(N)} \frac{1}{B_m} \sum_{j \in \mathrm{IND}(1)} E[\mathrm{BO}_{mj}(\bar{s})]. \tag{16}$$

As a consequence, maximising average availability is roughly equal to minimising the weighed average of expected backorders at each location.

As noted by Van Harten and Sleptchenko [22], the approximations (15) and (16) can be refined by using a second order approximation rather than a first order approximation. If we use that

$$(1-\varepsilon)^k \approx 1 - k\varepsilon + k(k-1)/2\,\varepsilon^2,$$

we can modify Eq. (16) as

$$A_m \approx 1 - \sum_{j \in \mathrm{IND}(1)} \left\{ \frac{E[\mathrm{BO}_{jm}(\bar{s})]}{B_m} - \frac{Z_j(Z_j-1)E[\mathrm{BO}_{jm}^2(\bar{s})]}{2B_m^2 Z_j^2} \right\} + \frac{1}{B_m^2} \sum_{i,j \in \mathrm{IND}(1) i<j} E[\mathrm{BO}_{im}(\bar{s})\mathrm{BO}_{jm}(\bar{s})]. \tag{17}$$

If the repair shops are modelled by M/G/∞ queues, the backorders of various items are mutually independent. If the multiplicity of each item Z_i equals one, which is not uncommon in practical situations (cf. [12]), the second term in the first summation drops out. Then we only need the mean number of backorders for all assemblies at all downstream locations to optimise availability.

However, the backorder distributions are *not* independent if they share the same repair facility having finite capacity. Then it is clear from formula (17) that we also need the correlations between backorders. In a multi-echelon, multi-indenture setting, these correlations may propagate throughout the network. For example, suppose that a central repair shop (say location 1) has finite capacity and repairs two different subassembly types (two classes), A and B. If subassembly A is part of assembly i and subassembly B is part of assembly j, there is theoretically a correlation between the backorders of assembly i at location $m \in \mathrm{CUS}(1)$ and the backorders of assembly j at another location $m \in \mathrm{CUS}(1)$. So the system availability at various locations (say A_n and A_m) may be dependent. However, this is not an issue when calculating the average availability A as in (16).

The optimisation procedure depends on the availability function and on its approximation. In principle, we can use the second order approximation, because the approximation for the multi-class M/M/k queue facilitates the computation of correlations between the pipeline distributions and hence also the computation of correlations between backorders. Van Harten and Sleptchenko [22] show the impact of using a second order approximation in a single-indenture, single-site model. Extension to general multi-indenture, multi-echelon models requires additional research.

Because we will base our numerical experiments on the case of the Royal Netherlands Navy as introduced by Rustenburg [12], we will use the alternative availability function (3). Therefore, our goal function for our computations in Section 6 is defined by (4). However, we emphasize that our approach also works with other definitions of availability.

6. Computational results

To test our model, we proceed as follows. We use the modified greedy optimisation heuristic as described in Section 5 to calculate stock levels for a range of experiments. We focus on cluster-dedicated repair shops (modelled by multi-class, multi-server queuing systems), because this is a new aspect in the spare parts management literature that has not been examined before. Each time, we choose the budget such, that 95% system availability can be reached. Then, we simulate the system to estimate the true availability. Also, we approximate the system availability using the traditional VARI-METRIC method, assuming that the repair shop has infinite capacity and by plugging in the observed repair shop throughput time in the $M/G/\infty$ model for the repair shop. Next, we examine to which extend the inclusion of finite repair capacity in the VARI-METRIC model improves availability estimates. In this approach we use the theory as sketched in Section 5.1 for deriving first and second moments of backorders at the operational level. Using these two moments, we fit a discrete distribution and find the probability of backorders at the operational level given a stock distribution. These backorder probabilities are the input of the greedy algorithm of Sherbrooke [4] that optimises the stock levels such, that the inventory investment to attain a given fixed availability (say 95%) is minimised. Here the system availability is calculated by (3).

In the experimental design, we focus on cases where we expect that finite capacity has significant impact. In Section 4, we already showed that the impact of finite capacity increases with repair shop utilisation and decreases with the number of servers. Therefore, we choose relatively high repair shop utilisation in our experimental design ($\geqslant 0.8$). We vary the following system characteristics: indenture structure, echelon structure, system failure rate, number of servers and repair shop utilisation rate. Altogether, we have 5 experimental factors. We will define two values for each factor and we will include all parameter combinations in our experimental design, resulting in $2^5 = 32$ experiments.

We use the system structure and the indenture structure as shown in Fig. 1, where numbers display probabilities to be sent to higher echelon repair r_{mj} and the cause probabilities q_{mjk}. For simplicity these probabilities are the same throughout the whole system.

We include both a two-echelon and a three-echelon system in our numerical experiment. Fig. 1 shows that the three-echelon system consists of 1 central depot, 4 frigates and 16 submarines. The two-echelon system does not have an intermediate frigate-level and simply consists of the depot and 16 submarines. Similarly, we analyse a three-indenture system as shown in Fig. 1 and the corresponding two-indenture system that arises when the lowest indenture level is omitted.

The installed base consists of one pump A and one pump B per submarine. We fix the failure rate of pump B in all experiments and vary the failure rate of pump A as shown in Table 1.

All locations have finite capacity cluster-dedicated repair shops. As clusters, we take all items at the same indenture level. That is, each location has two, respectively, three (multi-class, multi-server) repair shops in the two-, respectively, three-indenture model. One repair shop is dedicated to

Table 1
Values for the failure rates in the experiments (mean number of failures per pump per year)

Notation	Pump A	Pump B
Low	10	15
High	19	15

Table 2
Varying parameters of experiments

Echelons	Indentures	Failure rates	Number of servers per repair shop	Utilisation rates (%)
Three	Three	Low	3	80
Two	Two	High	10	95

pump A and pump B, one repair shop is dedicated to the valve, the piston and the flange and the third repair shop (only in the three-indenture model) is dedicated to the other items.

Next we have to specify the repair shop parameters, namely the mean repair times and the number of servers per repair shop. We choose these parameters such that the utilisation rates of all repair shops equal 80% and 95%, respectively. For the numbers of servers we define two different sets with 10 servers at each repair shop and with 3 servers, and repair times are defined completely by these parameters (arrival rates, utilisation rates and number of servers).

All parameter combinations (Table 2) lead to $2^5 = 32$ experiments. As mentioned above, we used our optimisation program to calculate the 32 different stock levels sets, which guarantee 95% system availability. Next, we estimated the true availability using discrete event simulation. The results of simulation runs (availability estimates with half-lengths of the 95% confidence intervals) are given below (Table 3).

It is clear that including the finite capacity of the repair shops leads to better results. The average deviation between simulation and the finite capacity approximation is 0.9%, whereas the average error is 4.8% if the infinite capacity assumption is used. Hence the inclusion of finite repair capacity improves the approximation accuracy by a factor 5.

To analyse these results in more detail, we examine how this approximation error depends on the five factors in our experimental design, see Table 4. The ordering of the factor levels in this table corresponds to Table 2.

This table shows us that the utilisation rates have the strongest effect on approximation error. The experiments with 80% utilisation rate have smaller approximation error than the experiments with 95% utilisation rate.

Repeating the same analysis for the VARI-METRIC method (see Table 5), we see that the average approximation error is considerably worse for all subsets of the experiments.

Table 4 shows that the utilisation of the repair facilities has the highest impact on the approximation error. We stress that our results are valid only within our experimental range. As we already showed in Section 4, the impact of the infinite capacity assumption is high if the repair shop utilisation is high, especially if the number of repair men in the shop is low (Fig. 3). Some side experiments showed that the infinite capacity model is appropriate indeed if the utilisation is low, especially if the number of servers is not too small. For example, the average error when using the infinite capacity assumption reduces to 1.5% if the utilisation equals 60% and the repair shop has 10 servers.

Heavy utilisation also leads to long queues and hence the simulation model requires more time to stabilise. This can be also seen in simulation data presented in van Harten and Sleptchenko [22] Table 2; where is shown that a single queue needs around 2000 subruns with each 1000 arrivals to arrive at 10% errors in estimation of mean number of items in queue. The fraction of failed items that are repaired at the lowest indenture repair shops is determined by the parameters r_{mj}. In our three-echelon model, we find that this fraction equals $0.2 \times 0.2 \times 0.2 = 0.008$ (or 0.04 in case of two-echelon models). A rough estimate of the simulation run length requirement for three-echelon models yields that we need 2000 (subruns) × 1000 (item failures per subrun)/0.008 = 250,000,000 failures at each submarine for each model (out of 32 models). We could not make such long runs and we stopped the simulations after 50,000–60,000 failures at each submarine. However, the errors *in the system availability* are within reasonable bounds. It shows that the errors in estimation of lowest indenture repair shops performance have only a limited impact on the estimation of the system performance. On the other hand, we see that the high utilisation of repair facilities

Table 3
Results of the approximation program and the simulation runs

Echelons	Indentures	No. of servers	Arrival rates	Utilisation rates (%)	Finite capacity (%)	Simulation	∞—Capacity approximation (%)
Three echelons	Three indentures	3	Low	80	95.03	93.9% ±0.39	99.89
				95	95.01	97.0% ±0.23	100.00
			High	80	95.06	94.8% ±0.29	99.92
				95	95.01	97.1% ±0.64	100.00
		10	Low	80	95.03	95.4% ±0.37	99.10
				95	95.01	96.9% ±0.53	100.00
			High	80	95.00	94.2% ±0.43	99.14
				95	95.00	96.9% ±0.45	100.00
	Two indentures	3	Low	80	95.02	94.6% ±0.33	99.91
				95	95.01	95.9% ±0.42	100.00
			High	80	95.01	95.9% ±0.71	99.86
				95	95.01	96.5% ±0.47	100.00
		10	Low	80	95.00	94.9% ±0.52	99.33
				95	95.01	95.7% ±0.48	100.00
			High	80	95.00	95.2% ±0.51	99.24
				95	95.01	96.1% ±0.54	100.00
Two echelons	Three indentures	3	Low	80	95.05	94.1% ±0.37	99.91
				95	95.01	95.3% ±0.52	100.00
			High	80	95.05	95.0% ±0.51	99.92
				95	95.01	96.9% ±0.41	100.00
		10	Low	80	95.01	94.0% ±0.50	99.92
				95	95.01	94.8% ±0.51	100.00
			High	80	95.05	95.3% ±0.45	99.91
				95	95.02	96.2% ±0.43	100.00
	Two indentures	3	Low	80	95.03	95.3% ±0.35	99.90
				95	95.01	96.2% ±0.56	100.00
			High	80	95.03	94.3% ±0.28	99.88
				95	95.01	96.0% ±0.51	100.00
		10	Low	80	95.02	94.7% ±0.42	99.37
				95	95.01	93.5% ±0.51	100.00
			High	80	95.00	95.3% ±0.45	99.24
				95	95.01	96.2% ±0.63	100.00

can destabilise the system. Despite the limit on the number of submarine failures, we still needed up to 200 hours for a single simulation run in some cases (three-echelon models). Although this is long, it is still manageable if the number of experiments is not too high as we did (32 experiments).

These experiments also show us that the stock levels at lowest echelon must be high to attain the target availability level (cf. Table 6). As a result, it can be cheaper to increase the capacity of the repair facilities then to keep many spare parts in stock. Hence, the optimisation of the availability with finite capacity repair shops can

Table 4
Average errors for all experiments and for different subsets of experiments

Average error	Echelons	Indentures	Failure rates	Number of servers	Utilisation rates
0.89%	3—1.01%	3—1.02%	Low—0.83%	3 set—0.97%	80%—0.50%
	2—0.77%	2—0.77%	High—0.96%	10 set—0.81%	95%—1.28%

Table 5
Average errors for the same set of experiments in case when availability was estimated with assumption of infinite repair capacity[a]

Average error	Echelons	Indentures	Failure rates	Number of servers	Utilisation rates
4.81%	3—4.76%	3—4.83%	Low—4.82%	3—4.93%	80%—4.63%
	2—4.86%	2—4.78%	High—4.80%	10—4.69%	95%—4.99%

[a] In fact we calculate the mean number of items in repair using formulas of multi-server queue, but the variance of this number we set up equal to the mean.

Table 6
Optimal stock levels at *each* location of a three-echelon three-indenture model with 10 servers and "high" failures rate using the finite capacity model

Utilisation	Echelons	Pump A	Pump B	Valve	Flange	Piston	Stem	Gasket	Rod	Ring
80%	1st echelon	2	3	7	13	16	4	4	5	7
	2nd echelon	3	5	6	12	16	4	3	4	6
	3rd echelon	12	13	8	15	19	5	4	5	7
95%	1st echelon	5	5	21	35	47	11	7	19	18
	2nd echelon	7	7	20	33	45	8	6	16	16
	3rd echelon	49	42	27	43	63	14	8	21	20

include a trade-off between stocks and servers capacities. A formal method for this trade-off still has to be developed and is a subject for further research.

Another interesting question is whether the use of our model with finite capacities leads to significantly different decisions than the traditional infinite capacity model. Therefore, we compared the stock allocations using both models. As an example, we show in Table 7 the stock allocation based on the traditional infinite capacity model for the same cases as in Table 6. As throughput times in the infinite capacity repair shops, we used the throughput times corresponding to repair shops with 10 servers and a utilisation of 80% and 95% in the finite capacity model, respectively.

The shift in stock allocation as shown in both tables are representative for our experiments. We see that decisions are modified in the following direction:

1. The total number of items on stock is considerable higher using the finite capacity model. As we also see from Tables 3–7, the infinite capacity model overestimates the real system availability. Therefore, the model suggests to put considerable less items on stock than actually required to reach the target availability.
2. Using the finite capacity model, a somewhat larger fraction of the stocks is allocated to the downstream locations.
3. There is no clear shift in stock distribution over indenture levels. That is, the distribution of stocks over first, second and third indenture is approximately similar for both the finite and infinite capacity model in our experiments.

Table 7
Optimal stock levels at *each* location of three-echelon three-indenture model with "high" failure rates using the infinite capacity model; the mean repair throughput times are equal to finite capacity repair shops with 10 servers and a utilisation of 80% and 95%, respectively

Mean repair throughput time	Echelons	Pump A	Pump B	Valve	Flange	Piston	Stem	Gasket	Rod	Ring
Equal to repair shop with 10 servers and 80% utilisation	1st echelon	3	4	6	12	15	4	4	4	6
	2nd echelon	4	6	6	11	14	3	3	3	6
	3rd echelon	9	11	7	13	16	4	4	4	7
Equal to repair shop with 10 servers and 95% utilisation	1st echelon	10	10	14	23	35	10	6	11	13
	2nd echelon	13	13	14	23	34	9	6	10	12
	3rd echelon	22	22	15	25	37	11	7	11	13

7. Conclusions

In this article we studied multi-echelon, multi-indenture service parts supply systems with finite repair capacity. We considered both item-dedicated repair shops, modelled by M/G/c queuing systems, and cluster-dedicated repair shops, modelled by multi-class M/M/c queuing systems. Our finite capacity VARI-METRIC approach yields more accurate results than the traditional approach of modelling repair shop throughput times by M/G/∞ queues if the utilisation is high (say >0.7), and especially if the number of servers is low (say <4). In our numerical experiments, we found that the average absolute deviation between estimated and simulated availability is only 0.9% using our method whereas this average deviation is 4.8% if we use the traditional infinite capacity VARI-METRIC approach. The VARI-METRIC approach generally overestimates the system availability.

Regarding the spare part stocks distribution within the system, we see that our finite capacity model leads to some shift in stock distribution to downstream locations. The distribution over the indenture levels is hardly affected, however.

We note that our conclusions are only valid within our experimental range, i.e. if the repair shop utilisation is relatively high. Naturally, our finite capacity model converges to the traditional infinite capacity approach if the utilisation decreases. Therefore, our approach is especially useful for practical situations in which (some of) the repair shops have a high work load of urgent repair jobs. Obviously, there are also cases where the traditional infinite capacity model is sufficient. This is true if outsourcing of repair jobs against little additional costs and lead time is possible or if the repair shop has low priority jobs (e.g. preventive maintenance or special projects) causing that the utilisation for high priority repair jobs is relatively low.

Furthermore, our method facilitates what-if analysis with important design parameters as the number of servers in the repair shops and the assignment of items to repair shop clusters. Another topic for further research is refining the finite capacity model for the repair shop to multi-class, multi-server priority systems. Such a model could enable us to improve further on the system efficiency. The latter issues have to be elaborated further in our next research phase. Such analysis is not possible for sure using the traditional infinite capacity VARI-METRIC approach.

Acknowledgements

We thank two anonymous referees for their suggestions to add explanation on several critical modelling choices and to improve the structure of our paper.

References

[1] M.A. Cohen, Y.-S. Zheng, V. Agrawal, Service parts logistics: A benchmark analysis, IIE Transactions 29 (1997) 627–639.

[2] J.H.C.M. Verrijdt, Design and control of service part distribution systems, Ph.D. Thesis, Eindhoven University of Technology, The Netherlands, 1997.

[3] C.C. Sherbrooke, METRIC: Multi-echelon technique for recoverable item control, Operations Research 16 (1968) 122–141.

[4] C.C. Sherbrooke, Optimal Inventory Modeling of Systems: Multi-echelon Techniques, Wiley, New York, 1992.

[5] W.D. Rustenburg, G.J. van Houtum, W.H.M. Zijm, Spare parts management at complex technology-based organizations: An agenda for research, International Journal of Production Economics 71 (2001) 177–193.

[6] F.M. Slay, VARY-METRIC: an approach to modeling multi-echelon resupply when the demand process is Poisson with gamma prior, Report AF301-3, Logistic Management Institute, Washington, DC, 1984.

[7] V.D.R. Guide Jr., R. Srivastava, Reparable inventory theory: Models and applications, European Journal of Operational Research 102 (1997) 1–20.

[8] H.F.M. de Haas, J.H.C.M. Verrijdt, Target setting for the departments in an aircraft repairable item system, European Journal of Operational Research 99 (1997) 596–602.

[9] A. Diaz, M.C. Fu, Models for multi-echelon repairable item inventory systems with limited repair capacity, European Journal of Operational Research 97 (1997) 480–492.

[10] M.A. Cohen, Y.-S. Zheng, Y. Wang, Identifying opportunities for improving Teradyne service-parts logistics system, Interfaces 29 (4) (1999) 1–18.

[11] W.J. Kennedy, J.W. Patterson, L.D. Fredendall, An overview of recent literature on spare parts inventories, International Journal of Production Economics 76 (2002) 201–215.

[12] W.D. Rustenburg, A system approach to budget-constrained spare parts management, Ph.D. Thesis, University of Twente, Enschede, The Netherlands, 2000.

[13] D.F. Pyke, Priority repair and dispatch policies for repairable-item logistics systems, Naval Research Logistics 37 (1990) 1–30.

[14] D. Gross, J.F. Ince, Spares provisioning for repairable items: cyclic queues in light traffic, AIIE Transactions 10 (3) (1978) 307–314.

[15] D. Gross, D.R. Miller, R.M. Soland, A closed queuing network model for multi-echelon repairable item provisioning, IIE Transactions 15 (4) (1983) 344–352.

[16] S.C. Albright, A. Soni, Markovian multi-echelon repairable inventory system, Naval Research Logistics Quarterly 35 (1988) 49–61.

[17] A. Gupta, S.C. Albright, Steady-state approximations for multi-echelon multi-indentured repairable-item inventory system, European Journal of Operational Research 97 (3) (1992) 340–353.

[18] S.C. Albright, A. Gupta, Steady-state approximation of a multi-echelon multi-indentured repairable-item inventory system with a single repair facility, Naval Research Logistics 40 (4) (1993) 479–493.

[19] Z.M. Avsar, W.H.M. Zijm, Resource-constrained two-echelon inventory models for repairable item systems, Working Paper, University of Twente Enschede, The Netherlands, 2000.

[20] J.-S. Kim, K.-C. Shin, S.-K. Park, An optimal algorithm for repairable-item inventory systems with depot spares, Journal of the Operational Research Society 51 (2000) 350–357.

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

[22] A. van Harten, A. Sleptchenko, On multi-class, multi-server queuing and spare part management, Working Paper, University of Twente, Enschede, The Netherlands, 2000, accepted for publication.

[23] I.J.B.F. Adan, M.J.A. van Eenige, J.A.C. Resing, Fitting discrete distributions on the first two moments, Probability in the Engineering and Informational Sciences 9 (1996) 623–632.

[24] J.H.A. de Smit, The queue GI/M/s with customers of different types or the queue GI/H_m/s, Advanced Applied Probabilities 15 (1983) 392–419.

[25] W. Whitt, Approximations for the GI/G/c queue, Production and Operations Management 2 (1993) 144–161.