Inventory reduction in spare part networks by selective throughput time reduction

M.C. van der Heijden *, E.M. Alvarez, J.M.J. Schutten

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

A R T I C L E   I N F O

Article history:
Received 15 September 2011
Accepted 12 March 2012
Available online 30 March 2012

Keywords:
Inventory
Spare parts
Repair time
Maintenance

A B S T R A C T

We consider combined inventory control and throughput time reduction in multi-echelon, multi-indenture spare part networks for system upkeep of capital goods. We construct a model in which standard throughput times (TPTs) for repair and transportation can be reduced at additional costs. We first estimate the marginal impact of TPT reduction on the system availability. Next, we develop an optimization heuristic for the cost trade-off between TPT reduction and spare part inventories. In a case study at Thales Netherlands with limited options for TPT reduction, we find a net saving of 5.6% on spare part inventories. In an extensive numerical experiment, we find a 20% cost reduction on average compared to standard spare part inventory optimization. TPT reductions downstream in the spare part supply chain appear to be the most effective.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Manufacturers of advanced capital goods such as computer systems and medical systems tend to expand their business by offering service contracts for system upkeep during the life cycle (Cohen et al., 2006). Such service contracts typically contain quantified service levels, such as a maximum response time in the case of a failure or a minimum uptime per year. We encountered such contracts at Thales Netherlands, a supplier of naval radar and combat management systems.

At the start of the contract, the supplier and/or the user invests in spare parts to facilitate fast repair by replacement of failed modules, the so-called Line Replaceable Units (LRUs). These (expensive) LRUs are often repaired rather than scrapped. Repair usually consists of diagnosis and replacement of a failed subcomponent, commonly referred to as Shop Replaceable Units (SRUs). Lack of spare SRUs leads to delay in LRU repairs, which increases the need for spare part inventories. Therefore, there is a trade-off between stocking LRUs and (cheaper) SRUs. Possibly, some SRUs are repairable themselves by replacing cheaper parts. So, we have a so-called multi-indenture product structure, see Fig. 1. We should decide about the stock levels of all items at all levels in the multi-indenture structure. In the remainder of this paper, we will use the phrases parent and child to refer to the relations in the multi-indenture structure: in Fig. 1, the supply cabinet is the parent of the power supply, and the power supply and air conditioning assembly are children of the supply cabinet. We use the term item for components at any level in the multi-indenture structure (LRUs, SRUs, parts).

Because the installed base is usually geographically dispersed, spare parts may be stocked at various locations. Stocks close to the sites where systems are installed are important for fast supply in case of a failure. This leads to several local stockpoints, each dedicated to a certain geographical area containing a part of the installed base. On the other hand, we may need central spare part stocks to take advantage of the risk pooling effect. Therefore, spare part supply systems usually have a multi-echelon structure as shown in Fig. 2. This is an example derived from a case study at Thales Netherlands, where we considered naval radars that are installed onboard of frigates. Spare parts may be stocked onboard, at the shore organization (close to a harbor), or at Thales Netherlands. In the remainder of this paper, we will use the common term base for a site where one or more systems are operational. We will use the phrases supplier and customer for the relations in the multi-echelon structure. In Fig. 2, Thales is the supplier of the Shore, and the Shore is a customer of Thales. Ready-for-use items are moved from the upstream part of the service supply chain (Thales) to the downstream part (ships).

To optimize the initial spare part inventories, Thales uses a commercial tool based on the VARI-METRIC method (Sherbrooke, 2004). If there is evidence during contract execution that the actual service levels are below target (e.g. in terms of downtime waiting for spare parts), the service provider intervenes. At a tactical level, options are among others (i) buying additional spare parts, (ii) reducing repair shop throughput times, and (iii) reducing transportation times of spare parts. In this research, we focus
on throughput time (TPT) reduction (of repair and transportation) as alternatives to spare part investment for multi-indenture, multi-echelon spare part networks. At Thales Netherlands, such reductions are feasible at extra costs. It is well known that influencing repair TPT for specific items may have a large impact on the total costs (Sleptchenko et al., 2005; Adan et al., 2009).

To gain insight on the impact of TPT reductions, we first develop expressions for the marginal backorder reduction of LRUs at operating sites as a function of the marginal reduction in TPT of each repair and transport in the network. We use these expected backorders as criterion, because their minimization is approximately equivalent to maximizing operational availability (Sherbrooke, 2004). If pipelines are Poisson distributed, we need only the fill rates of all items in the multi-indenture structure at all locations in the multi-echelon networks for this purpose. Combining these marginal values with a certain discrete step size for the TPT reductions, we develop a heuristic optimization method to balance the investment in TPT reductions to investment in extra spares. In a numerical experiment, we show that a trade-off between spare part inventories and TPT reductions may yield considerable cost savings.

In this paper, we first discuss related literature and state our contribution (Section 2). We define our model in Section 3. Section 4 shows how we estimate the impact of TPT reduction for given spare part stock levels. This is input for our optimization heuristic (Section 5). In Section 6, we discuss numerical results from both the case study at Thales Netherlands and a large set of theoretical problem instances. We end up with conclusions and directions for further research in Section 7.

2. Literature

There is a vast extent of literature on optimization of slow moving spare part inventories in multi-echelon, multi-indenture supply chains (Sherbrooke, 2004; Muckstadt, 2005). These models contain many parameters, some of them resulting from underlying decisions. Examples are the location and allocation of repair activities, repair and transportation lead times, and item failure rates. In the last decades, several models have been developed that consider some of these decisions jointly. Oner et al. (2010) consider the joint decision of mean time between failures (which can be influenced during product design) and the costs of spare parts during the life cycle for a single item. Joint decisions for spare parts inventories and repair locations, taking into account the costs of resources required, are discussed by among others Alfredsson (1997) and Basten et al. (under review a). Rappold and Van Roo (2009) combine the spare part stocking problem with facility location. Focusing on the relation between spare part inventories and TPTs, there are two streams of literature:

- analysis and optimization of spare parts and repair and supply processes at a tactical level, where a selected subset of items is given high priority in repair;
- operational optimization of spare part networks by dynamic priority setting in repair and supply, given fixed spare part stock levels and resource capacities.

Within the stream focusing on the tactical level, we distinguish the selective use of emergency repair and supply in case of low stocks, and priority setting models with finite repair capacities. In the first area, Verrijdt et al. (1998) use a single item model to show the impact of emergency repairs if the stock level drops below a certain threshold value. Perlman et al. (2001) consider a single-item, two-echelon model with finite capacity repair shops and assume that emergency repair is applied with a certain probability. Van Utterbeeck et al. (2009), on the other hand, focus on supply flexibility, i.e., the performance improvement if emergency shipments and lateral transshipments are allowed. They use simulation optimization to search the optimal system design and stock allocation, again for a single item.

The models with finite repair capacities usually model the repair shops as single or multi-server queues with exponentially distributed repair times (Gross et al., 1983; Diaz and Fu, 1997; Sleptchenko et al., 2003). An important issue in this line of research is the trade-off between repair capacity and spare part inventories: limited capacity leads via high utilization and long TPTs to more spare part stocks. Sleptchenko et al. (2005) introduce priority queueing models for the repair shop where the items are assigned to two priority groups (high or low priority). They show that appropriate priority assignment may lead to a significant reduction in the spare part inventory investment. The idea is to prioritize repair of items with high value and small repair times, so that the work-in-process of these items is reduced with limited impact on other items. A similar idea has been used by Adan et al. (2009), who consider multiple priority classes (> 2) in a single-location, single-indenture problem. They develop a method for exact cost evaluation.

At the operational level, various priority rules have been examined. These models assume that all resources are given (spare part inventories, repair capacities) and search for efficiency gain using (i) repair priorities (if a server becomes idle, which item from the queue should be repaired first?), and (ii) dispatch priorities (if an item has been repaired and there are multiple outstanding orders for this
item, which order should be filled first?). Regarding repair priorities, Hausman and Scudder (1982) discuss several rules in a single-location, three-indenture model. The best rules lead to a backorder reduction equivalent to 20% less inventories. Hausman (1984) extends this model to the multiple failure case and finds similar results. Pyke (1990) combines repair priorities with dispatch policies in a simulation study and concludes that priority repair improves the system performance, whereas dispatching priorities have limited impact. Caggiano et al. (2006) develop two methods to set repair and dispatch priorities in two-echelon networks within a finite planning horizon. They show that significant gains are feasible in a rolling horizon setting. Tiemessen and Van Houtum (2010) show that operational priorities may yield about 10% cost reduction on top of static repair priorities in a multi-item, single-location model.

The focus in our paper is on the impact of repair and transportation time differentiation at a tactical level. Inspired by the Thales case, we aim for a realistic model, i.e., a multi-item, multi-indenture, multi-echelon setting, whereas many papers address single items models that can be used as a building block only. In contrast to the work on finite capacity models, we do not model the repair shops by finite capacity (multi-server) queues for the following reasons. First of all, repair shops often have more similarity to a job shop environment that could be modeled as a queueing network rather than a multi-server queue. Further, repair capacities are often not fixed or may be fuzzy, because a repair shop may have other tasks than spare part repair only. Also, flexibility options such as working overtime or temporarily hiring personnel may exist. If repair is outsourced, the repair capacity is even unknown, and the repair lead times and corresponding prices are the result of a negotiation process. Therefore, we choose a model in which we may select different options for repair and transportation lead time at different prices, without explicit capacity modeling. We encountered this situation at Thales, which offers both a normal repair and a fast repair option to its customers without service contracts at different prices. The same flexibility could be used to optimize the performance for customers having service contracts. This also holds for emergency transportation that Thales could apply for certain combinations of items and locations against additional costs.

Summarized, we aim to contribute the following to the literature:

1. We consider a simple but practical model for the trade-off between spare part stocks and TPT reduction in repair and transportation, based on pricing of TPT reduction. This model is suitable for multi-item, multi-echelon, multi-indenture networks as we encountered at Thales Netherlands.

2. We use estimates for the marginal impact of TPT reductions to develop an efficient heuristic for the simultaneous optimization of spare part inventories and repair and transportation TPTs. We show that significant cost reductions are feasible (20% on average for theoretical problem instances).

3. We show how the savings depend on type of problem instance and we characterize the type of policies that we typically find. In particular, we see that TPT reductions are most profitable downstream in the network.

4. We apply our method in a case study at Thales Netherlands and find interesting savings (5.6% on the inventory investment). The restricted options for reduction of TPTs downstream in the network cause lower savings than in the theoretical experiments.

3. Model, assumptions, and notation

We consider a multi-indenture, multi-echelon spare part network. Our decision variables are spare part inventory levels, and repair and transportation TPTs of all items at all locations in the network. For each combination of item and location, we have a discrete set of TPTs, and costs are attached to each option.

3.1. Assumptions

We proceed from the standard assumptions of the VARI-METRIC model (Sherbrooke, 2004):

1. Systems fail according to a stationary Poisson process.
2. All failures are critical, i.e., they cause system downtime.
3. Each item failure is caused by the failure of at most one subcomponent.
4. Repair shops are modeled as \( M/G/\infty \) queues, where successive repair TPTs of the same item at the same location are independent and identically distributed.
5. The flow of repair jobs of each item arriving at each location is given. This is modeled as a given fraction of jobs that can be repaired (the rest is forwarded for repair upstream).
6. All items are as good as new after repair.
7. Requests for spare parts are handled First Come, First Serve (FCFS).
8. We use an \((s-1, s)\) one-for-one replenishment policy for all items at all locations.
9. Any customer location has one unique supplier (except the most upstream stockpoint).
10. Inventories are always replenished from the direct supplier in the multi-echelon structure, i.e., there is no lateral supply between locations at the same echelon.
11. All transportation TPT (or: order-and-ship times) are deterministic.

With respect to TPTs (repair and transportation), we further assume:

12. For each combination of item and location, we have a discrete set of TPTs, and costs are attached to each option. This corresponds to the practice at Thales Netherlands, where a limited set of options were available for both repair and transportation TPT (see the case description in Section 6.3).

With respect to the latter assumption, we proceed from a standard repair and transportation lead time for each combination of item and location, and we consider options for TPT reductions that we may select at additional costs. Without loss of generality, the additional costs per repair are increasing in the repair TPT reduction, and the same applies to the transportation costs. If not, we ignore inferior options.

3.2. Notation

We use similar notation as in Sherbrooke (2004) and distinguish input parameters, decision variables, auxiliary variables, and performance measurements (output).

3.2.1. Input

\[ J \] set of all locations in the network
\[ B \] set of all bases, i.e., all locations in the network where systems are installed
\[ I \] set of all items
\[ L \] set of all LRUs, i.e., all first indenture items
\[ m_{ij} \] demand rate for item \( i \) at location \( j \) \((i \in I, j \in J)\)
\[ r_{ij} \] fraction of demand for item \( i \) at location \( j \) that can be repaired at the same location (the rest is forwarded to the supplier of \( j \) for repair)
fraction of item $k$ failures that is due to a failure of item $i$ costs per year for holding one item $i$, these costs may include costs of capital, storage and risk, including the obsolescence risk $T_i$ \( n \)th option for the repair shop TPT of item $i$ at location $j$, which is strictly decreasing in $n$; index $n=0$ gives the standard repair throughput time $O_j$ \( n \)th option for the transportation TPT of item $i$ at location $j$, which is strictly decreasing in $n$; index $n=0$ gives the standard transportation time $C^R_j$ costs per repair if the repair shop TPT of item $i$ at location $j$ equals $t$ $C^M_j$ costs to move a single item $i$ to location $j$ from its supplier if the transportation TPT equals $t$.

Note that the demand rates $m_{ij}$ are input for all LRUs $i \in L$ and all bases $j \in B$. We can recursively find the demand rates for all other combinations of item $i$ and location $j$ from $m_{ij} = m_j q_{ij} + \sum_{c \in D_j} m_{ck} (1-r_{ck})$, where $k$ is the parent of $i$ and $D_j$ denotes the set of all customers of location $j$.

3.2.2. Decision variables $s_{ij}$ inventory level for item $i$ at location $j$ $a_{ij}$ index of repair TPT of item $i$ at location $j$ $b_{ij}$ index of transportation TPT of item $i$ at location $j$.

We denote the matrices of decision variables for all items and all locations in bold face by $\mathbf{s}$, $\mathbf{a}$ and $\mathbf{b}$.

3.2.3. Auxiliary variables $f_j(n)$ probability that the number of items $i$ in the pipeline at location $j$, i.e., all items in repair or in resupply, equals $n$; we denote the corresponding mean by $\mu_j$. $f_j(s,a,b)$ Expected backorders of item $i$ at location $j$ under policy $(\mathbf{s}, \mathbf{a}, \mathbf{b})$.

3.2.4. Performance measurement $\beta_j(s,a,b)$ Fill rate of item $i$ at location $j$ under policy $(\mathbf{s}, \mathbf{a}, \mathbf{b})$, i.e. the fraction of demand that can be filled from stock on shelf without delay.

3.3. Model

As in VARI-METRIC, we aim to balance the operational availability and the costs required for holding spare part inventories, and, in our case, the costs of repair and transportation. As mentioned before, Sherbrooke (2004) uses the sum of the backorders of LRUs at bases (sites where systems are installed) as a proxy for the operational availability. Therefore, we find the following nonlinear optimization model:

\[
\begin{align*}
\min_{\mathbf{s}, \mathbf{a}, \mathbf{b}} & \quad \sum_{i \in L, j \in B} (s_{ij} + m_{ij} q_{ij} C_j^R(T_j(a_{ij})) + m_{ij} (1-r_{ij}) C_j^M(O_j(b_{ij}))) \\
\text{subject to} & \quad \sum_{i \in L, j \in B} EBO_j(\mathbf{s}, \mathbf{a}, \mathbf{b}) \leq EBO^{\text{target}}
\end{align*}
\]

where $EBO^{\text{target}}$ denotes a target number of LRU backorders at bases corresponding to a certain operational availability. We can interpret this target backorder sum as the average number of systems that are down waiting for a spare part. We can set this target as $EBO_{\text{target}} = (1-\text{availability}) \cdot IB$, where $IB$ denotes the total number of systems in the installed base. The expected backorders of item $i$ at location $j$ are computed by

\[
\begin{align*}
EBO_j(s,a,b) = \sum_{n=s_j+1}^{\infty} (n-s_j) f_j(n|\mathbf{s}, \mathbf{a}, \mathbf{b})
\end{align*}
\]

with the probability density of the pipeline $f_j(n|\mathbf{s}, \mathbf{a}, \mathbf{b})$ depends on the repair and transportation TPTs of item $i$ at location $j$, and the probability distribution of the backorders, or the probability distribution of the availability $\beta_j(s,a,b)$, i.e. the probability that item $i$ is available at location $j$ and the availability of all children downwards in the multi-indenture structure, at location $j$ and all locations upstream in the supply chain. The same applies to the repair shop and transportation TPTs and so for the impact of the decision variables $d_{ij}$ and $b_{ij}$. In METRIC, all pipeline distributions were originally approximated by Poisson distributions. Because this approximation can be quite bad, two-moment approximations for the pipelines have been used in VARI-METRIC (Sherbrooke, 2004). This can be done using negative binomial distributions, because the variance-to-mean ratio of the pipelines are usually $\geq 1$. As a more general solution, we use the method of Adan et al. (1995) to fit a discrete probability distribution function to the first two moments. Hence, we compute an approximation of all backorders using two-moment approximations for the pipeline distributions.

VARI-METRIC considers only the stock levels and not the TPT reduction. For optimization, a simple greedy heuristic is typically applied. That is, starting at all stock levels $s_j=0$, we add in each iteration an item of type $i$ to stock at a certain location $j$ that has the largest ratio of reduction of expected backorders of LRUs at bases and the additional inventory investment. In popular terms, this heuristic is referred to as the biggest bang for the buck.

Problem (P1) is a large nonlinear integer programming problem having three times as many decision variables as VARI-METRIC: next to the stock levels $s_{ij}$, we also have to decide about the repair and transportation TPTs for all combinations of item $i$ and location $j$. As VARI-METRIC is an optimization heuristic, it is reasonable to expect that P1 cannot be solved exactly in a reasonable amount of time for problem instances with a realistic size. So, we focus on optimization heuristics.

4. Analysis of TPT reduction

We first specify the impact of TPT reduction on the expected backorders of LRUs at the bases. Under the assumption of Poisson distributed pipelines, we find the partial derivatives of the total expected backorders of LRUs at bases to any mean repair TPT and order-and-ship time in the following way. Then, the density function of the pipeline $f_j(n|\mathbf{s}, \mathbf{a}, \mathbf{b})$ is given by

\[
f_j(n|\mathbf{s}, \mathbf{a}, \mathbf{b}) = \frac{\mu_j^m e^{-\mu_j}}{n!}
\]

where the mean pipeline $\mu_j$ depends on the decision variables $(\mathbf{s}, \mathbf{a}, \mathbf{b})$. From here on, we will use the shorthand notation $(\cdot)$ for a variable if $\cdot$ is a function of some of the decision variables $(\mathbf{s}, \mathbf{a}, \mathbf{b})$.

Using elementary calculus, we can derive from (1) and (2) that

\[
\frac{\partial EBO_j(\cdot)}{\partial T_j} = \sum_{n=s_j}^{\infty} \frac{\mu_j^m e^{-\mu_j}}{n!}
\]

which equals $1-\beta_j(\cdot)$, i.e. one minus the fill rate. For a single site model, we have that $\mu_j = m_j T_j$, and so we find using the chain rule
In a two-echelon, single-indenture model with location 0 as the supplier of location $j$, we have (Sherbrooke, 2004):

$$\frac{\partial EBO_i}{\partial T_{ij}} = \frac{\partial EBO_i}{\partial \mu_j} \left(1 - \beta_j(\cdot)\right) \mu_j$$

(4)

In a two-echelon, single-indenture model with location 0 as the supplier of location $j$, we will see that we do not use the exact values of the moment approximations for the pipelines. In the next section, however, we will see that we do not use the exact values of the partial derivatives, but use only their ranking to select the most promising option (repair or shipment) for TPT reduction.

5. Optimization heuristic

At first sight, we can easily extend the greedy heuristic for spare part optimization by adding extra options for TPT reduction. We estimate the impact of repair TPT reduction of item $i$ at location $j$ on the total LRU backorders using the partial derivatives as found in the previous section: $T_{ij}(a_j) - T_{ij}(a_j + 1)\sum_{k \neq i} \sum_{b \neq j} \beta_{bij}$. This is obviously an approximation, but it gives us a good idea on the impact of TPT reductions. We compare this impact to the additional costs, they being additional repair costs times the number of repairs per year: $(C_{ij}(T_{ij}(a_j + 1)) - C_{ij}(T_{ij}(a_j))) \mu_j \beta_j$. So, we have the following simple approximation for backorder reduction per euro $\Delta_b(a_j)$ due to repair TPT reduction:

$$\Delta_b(a_j) = \frac{T_{ij}(a_j) - T_{ij}(a_j + 1)}{(C_{ij}(T_{ij}(a_j + 1)) - C_{ij}(T_{ij}(a_j))) \mu_j \beta_j}$$

(9)

The backorder reduction per euro due to transportation TPT reduction $\Delta_b(b_j)$ equals

$$\Delta_b(b_j) = \frac{O_{ij}(b_j) - O_{ij}(b_j + 1)}{(C_{ij}(O_{ij}(b_j + 1)) - C_{ij}(O_{ij}(b_j))) \mu_j (1 - \beta_j)}$$

(10)

We denote the standard backorder reduction per euro from VARI-METRIC, due to adding a spare part $i$ at location $j$ to stock, by $\Delta_b(s_j)$. Now a logical extension of the greedy VARI-METRIC heuristic is to add all options for TPT reduction, and to select at each iteration the decision that yields the highest backorder reduction per euro spent. This can be either adding a spare part to stock, or a discrete step reduction in either repair or transportation TPT. Unfortunately, this heuristic does not work well, since TPTs and stock levels are not independent: if we add stocks, the impact of TPT reduction decreases. We typically see that we initially decide to reduce many TPTs, because there are hardly any spare part stocks and so the impact of TPT reduction is high. If spare part stock levels are zero, any hour reduction of TPT is an hour reduction in system downtime. When we have added spare parts to stock, we find out that the impact of these TPT reductions decreases, and finally we may even end up with a solution that is worse than VARI-METRIC, ignoring the options for TPT reduction. So, we have to find another heuristic.

As the problems above are caused by the generally decreasing impact of TPT reduction on the spare part inventories, it seems better to construct a heuristic that considers TPT reduction while stock levels are decreasing rather than increasing. The basic idea is the following: first, we apply VARI-METRIC using the standard TPTs $T_{ij}(0)$ and $O_{ij}(0)$. Then, we improve this solution by reducing TPTs and stock levels are not independent: if we add stocks, the impact of TPT reduction is high. If spare part stock levels are zero, any hour reduction of TPT is an hour reduction in system downtime. When we have added spare parts to stock, we find out that the impact of these TPT reductions decreases, and finally we may even end up with a solution that is worse than VARI-METRIC, ignoring the options for TPT reduction. So, we have to find another heuristic.

As the problems above are caused by the generally decreasing impact of TPT reduction on the spare part inventories, it seems better to construct a heuristic that considers TPT reduction while stock levels are decreasing rather than increasing. The basic idea is the following: first, we apply VARI-METRIC using the standard TPTs $T_{ij}(0)$ and $O_{ij}(0)$. Then, we improve this solution by reducing the spare part levels having the least added value for TPT reductions having the most added value. The spare part having

![Fig. 3. Computation scheme for the partial derivatives of LRU backorders at bases.](image-url)
the least added value is the last one we added to stock in the VARI-METRIC algorithm. We search the best (set of) TPT reduction(s) compensating the loss of availability by removing the latter spare parts. If these TPT reductions cost less per year than the holding cost of the removed spare part, we accept the stock level reduction. We continue until no improvement is found. So, our basic algorithm is as follows.

Basic optimization heuristic

1) Initialize the decision variables: \( s_{ij} = 0, a_{ij} = 0, b_{ij} = 0 \) (\( i \in I, j \in J \)).

2) Use VARI-METRIC to optimize the spare part stock levels for the TPTs \( T_{ij}(a_{ij}) \) and \( O_j(b_{ij}) \) (\( i \in I, j \in J \)). Keep track of the order in which spare parts are added to stock (item \( i \), location \( j \)). Let us denote that list by \( (i_n, j_n) \), this being the type of item \( i_n \) and the location \( j_n \) that has been added to stock in iteration \( n (n = 1..N) \). Compute all partial derivatives \( P_{ij} \) and \( Q_{ij} \).

3) Consider compensating stock reduction of spare part \((i_n, j_n)\) by TPT reduction. The cost savings per year are \( h_{in} \). Set the additional costs for TPT reduction equal to \( C_{TR} = 0 \). Set \( i^* = i_n \) and \( j^* = j_n \).

a. Recompute the expected backorders and the partial derivatives that have changed, that is, for all combinations of (i) items in the same branch of the multi-indenture structure as \( i^* \) (parents and children), and (ii) locations in the same branch of the multi-echelon structure as \( j^* \) (customers and suppliers). If the sum of expected LRU backorders at bases is greater than or equal to the target EBO\(^{upper}\) then go to Step 3b else go to 3c.\(^1\)

b. Select the best TPT reduction from the options \( a_{ib}, b_{ij} \) by selecting \((i^* j^*)\) from

\[
(i^* j^*) = \arg\min_{i,j} \{\min(\Delta h_n(a_{ib}), \Delta h_n(b_{ij}))\}
\]

If the minimum is attained for a repair TPT reduction, then set

\[
C_{TR} = C_{TR} + \left\{ C_{TR}(T_{tr}^*(a_{tr}^* + 1)) - C_{TR}(T_{tr}^*(a_{tr}^*)) \right\} m_{tr} r_{tr}^* \]

and

\[
a_{tr}^* = a_{tr}^* + 1 \]

else set

\[
C_{TR} = C_{TR} + \left\{ C_{TR}(O_{tr}^*(b_{tr}^* + 1)) - C_{TR}(O_{tr}^*(b_{tr}^*)) \right\} m_{tr} (1 - r_{tr}^*) \]

and

\[
b_{tr}^* = b_{tr}^* + 1 \]

Return to step 3a.

c. If \( C_{TR} < h_{in} \), the costs of TPT reduction are less than the cost savings of removing a spare part, whereas we attain the target backorder level. Accept this stock reduction and go to Step 4. Otherwise, keep item \( i_n \) on stock at location \( j_n \) and STOP.

4) \( N := N - 1; \) If \( N \geq 1 \) and there are still options for TPT reduction left, then consider the next spare part for stock reduction: go to Step 3.

Because we have to update only a limited number of partial derivatives each time we modify spare part stock levels or TPTs (Step 3a), the algorithm is pretty fast (from a fraction of a second to various minutes, depending on the size of the problem). The basic heuristic stops if it is not cost effective to reduce TPT to compensate for stock reduction of a single spare part. A straightforward extension is to consider stock reduction of two or more spare parts simultaneously, compensated by pieces of TPT reduction. In principle, we can continue until we run out of either options for spare part reduction or options for TPT reductions, whichever come first (usually the TPT reductions come first). This may seriously increase the computation times, however. As a compromise, we consider stock reduction of multiple spare parts compensated by one or more pieces of TPT reduction, until the next best marginal effect of TPT reduction according to criterion (11) is less that the impact of removing the next spare part, this being the total increase in LRU backorders at the bases divided by the decrease in costs \( h_{in} \).

An obvious drawback of our heuristic is that the optimality gap is unknown. An optimal algorithm, however, is not easy to find. An option is an approach similar to the method by Basten et al. (under review a) for the integration of decisions for repair locations and resource locations (Level of Repair Analysis) and spare part inventories. Such an approach is out of scope for this paper (see also Section 7). Advantages of our heuristic are its simplicity and speed, such that we are able to analyze models of realistic size. Moreover, the construction of the heuristic guarantees that we find only solutions that are at least as good as the standard VARI-METRIC procedure without considering TPT reductions.

6. Experiment and results

In this section, we design a numerical experiment to analyze the savings that can be obtained using joint optimization of spare part inventories and TPTs and to characterize its type of policies. We give our experimental design in Section 6.1, and discuss our results in Section 6.2. We illustrate our method in a case study at Thales Netherlands (Section 6.3).

6.1. Experimental design

We focus on two-echelon, two-indenture networks. The holding cost rate is 25% of the item value per year. We vary the size and type of the problem as specified in Table 1.

For each setting, we generate randomly 25 problem instances as follows:

1) we draw the demand per year per base for each LRU \( m_{ij} \) (\( i \in I, j \in B \)) from a continuous uniform distribution around the mean with minimum demand rate 0.002;
2) we randomly assign the SRUs to LRUs using equal probabilities;
3) if an LRU has one or more SRUs, the probability that no SRU needs to be replaced upon LRU failure is always 0.1, whereas the remaining 0.9 probability mass is allocated to the SRUs based on a continuous uniform distribution (giving the cause probabilities \( q_{ia} \)).

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Experimental factors.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental factor</strong></td>
<td><strong>Low value</strong></td>
</tr>
<tr>
<td>Number of LRUs</td>
<td>25</td>
</tr>
<tr>
<td>Average number of SRUs per LRU</td>
<td>0.5</td>
</tr>
<tr>
<td>Average demand per LRU per base ( m_{ij} ) (per year)</td>
<td>0.05</td>
</tr>
<tr>
<td>Number of bases</td>
<td>3</td>
</tr>
<tr>
<td>Average repair time ( T_0 ) over all items (year)</td>
<td>0.05</td>
</tr>
<tr>
<td>Repair costs as a percentage of the item value (%)</td>
<td>15</td>
</tr>
<tr>
<td>Transportation costs (€)</td>
<td>100</td>
</tr>
<tr>
<td>Target availability</td>
<td>0.95</td>
</tr>
</tbody>
</table>

\(^1\) This will never occur in the first iteration, but may happen in next iterations.
4) we draw the net value per item from a shifted exponential distribution with lower bound $h_{400}$ and mean $h_{6000}$; the gross LRU value includes the net values of its SRUs.

5) all items can be repaired at the central depot ($r_{ij} = 1$ if $j$ represents the central depot). At the bases, the repair probabilities $r_{ij}$ depend only on the item $i$ and are drawn from a continuous uniform distribution on the interval $[0.1, 0.9]$.

In all cases, we consider the following options for TPT reduction (Table 2):

We use fewer options for transportation TPT reduction, because these times are usually much smaller than repair TPTs. Altogether, our experiment consists of $2^8$ (8 experimental factors) $\times$ 25 (random problem instances per setting) = 6400 problem instances.

6.2. Numerical results

6.2.1. Savings percentage

We compute the cost savings from including throughput reductions as decision variables in the optimization. That is, we compute the total costs as specified in the goal function of optimization problem (P1) in Section 3 after optimization to the total costs after Step 1 of our algorithm (i.e., application of VARI-METRIC using standard TPTs only). Over all 6400 problem instances, we find average cost savings of 19.8%.

Fig. 4 shows the impact of the experimental factors on the cost savings, sorted by magnitude of impact. We observe that the average demand per LRU has the highest impact: TPT reduction is particularly profitable if demand is low. This makes sense, because repair and transportation costs increase proportionally in the demand, whereas spare part holding costs increase less than proportionally because of the portfolio effect. Further, the savings percentage decreases with the target availability, the number of LRUs in the system, the mean repair costs, and the mean repair time. The impact of the average availability and the number of LRUs are remarkable. In both cases, the average downtime allowed per LRU decreases. A possible explanation is that low downtime requirements per LRU lead to high spare part stock levels, and then the impact of TPT reduction is relatively low. The other factors (average number of SRUs per LRU, number of operational sites, transportation costs) have a marginal impact on the cost savings. We expect that higher transportation costs would lead to less reduction in order and ship times and so to less cost savings. We do not see this in the savings percentage, but we see it in the type of policy that we choose. We will discuss these policies in more detail in Section 6.2.2.

Table 2

<table>
<thead>
<tr>
<th>Repair</th>
<th>Transportation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPT reduction</td>
<td>Cost increase</td>
</tr>
<tr>
<td>25%</td>
<td>40%</td>
</tr>
<tr>
<td>50%</td>
<td>100%</td>
</tr>
</tbody>
</table>

6.2.2. Type of policy

To examine the type of policy we find for the TPTs, we measure the degree of TPT reduction in a single problem instance by the weighted average percentage TPT reduction with the number of (repair or transportation) jobs as weights. We distinguish between the levels in the multi-echelon system and the levels in the multi-indenture structure. Obviously, we find most TPT reduction in the problem instances with the highest savings. Apart from that observation, the following observations are interesting:

- The average reduction in repair TPT is 8.5% for all upstream repairs and 24.8% for all downstream repairs. Clearly, we have most TPT reductions downstream in the network.
- We observe most TPT reduction for repairs downstream (at the bases) when repair costs are low and repair times are high (38% reduction).
- We hardly use repair TPT reduction of SRUs at the central depot (6.4% on average). We find the most reduction in the case of few bases and low demand rates (still only 12.7%).
- The average reduction in transportation TPT between central depot and bases is 25%.
- Although the transportation costs have little impact on the savings (see Fig. 4), they influence the type of policy. The average TPT reduction is 34% if the costs per shipment are $\varepsilon100$, and 16% for transportation costs of $\varepsilon500$. So, we indeed reduce the transportation TPT less if the costs are higher.

Fig. 4. Impact of the experimental factors on the average cost savings (see Table 1 for the high and low settings per factor).
6.2.3. Impact of scenarios for TPT reduction

If we consider only repair TPT reductions and no transportation TPT reductions, we still get significant average cost savings, namely 14.9% instead of 19.8%. If we limit the options to transportation TPT reductions however, the average cost savings are 3.3% only. It is remarkable that the joint effect of repair and transportation TPT reductions is larger than the sum of the separate effects.

Next, we analyze the impact of the number of scenarios for TPT reduction (i) by excluding scenarios for repair TPT reduction: we only allow cutting repair TPTs in only half at twice the costs (ii) by adding scenarios for TPT reduction. In the latter case, we considered the following options for both repair and transportation TPTs as given in Table 3. We find that the average savings decrease from 19.8% to 16.0% if we reduce the number of options for TPT reduction. Under additional options, the average savings increase from 19.8% to 22.3%. So, the number of discrete steps in TPT reduction has impact, but it is not very large. We already achieve significant gains with a single alternative option for TPTs only.

6.2.4. Three-echelon, three-indenture systems

To examine whether our findings remain valid for other network types, we designed a similar experiment for three-echelon, three-indenture networks. The cost savings are somewhat higher on average (24.8%), but the other findings are similar to two-echelon, two-indenture systems. The only new finding is that we observe a larger impact of the multi-indenture structure on the cost savings. Higher savings are feasible for the combination of more SRUs per LRU and more subcomponents per SRU, so for a “heavier” multi-indenture structure (29.2% savings). We particularly observe a higher reduction in repair TPTs as well as transportation TPTs downstream (and particularly for LRUs).

6.3. Case study

To evaluate our method in a practical setting, we collected data for a part of a radar system at Thales Netherlands. The detailed data are confidential, but we give an outline of the key characteristics below. The data are related to a service contract covering six radar systems onboard of six frigates. Spare parts are supplied in a three echelon system from Thales Netherlands via a shore organization to the frigates. Spare parts may be stocked and repaired at each of the three levels. The subsystem consists of 114 different items, spread over two indenture levels (LRUs and SRUs). The item values vary from a few hundreds of euros to more than €100,000 (LRU including SRUs). The options for TPT reduction are:

- Repairs at Thales Netherlands can be processed via a “fast channel” at extra labor costs, yielding a repair TPT reduction of 50% on standard values of several months. The extra labor costs are related to the product value and may easily exceed €1000.

Table 3 Scenarios for repair and transportation TPT reductions (%).

<table>
<thead>
<tr>
<th>Repair TPT reduction</th>
<th>Cost increase reduction</th>
<th>Transportation TPT reduction</th>
<th>Cost increase reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>25</td>
<td>40</td>
<td>25</td>
<td>40</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>60</td>
<td>300</td>
<td>60</td>
<td>300</td>
</tr>
<tr>
<td>75</td>
<td>700</td>
<td>75</td>
<td>700</td>
</tr>
</tbody>
</table>

Application of our heuristic yields 6.3% savings on the spare part holding costs at extra repair and transportation costs equal to 0.7% of the original inventory investment, so we have a net saving of 5.6%. Note that this is not a percentage over the total spare part holding, repair and transportation costs, since we were not able to specify repair and transportation costs for the standard TPTs. In fact, we only need the additional costs of TPT reduction only to apply our method. Although the savings are relevant for Thales Netherlands given the amount of money involved, it is clear that the savings are considerably less than the average that we observed in our theoretical experiments. We have the following explanation for this:

- The theoretical experiments show that TPT reduction downstream in the network is usually most profitable. However, Thales Netherlands can influence repair times only at the own site, since both the shore and the ships are part of the customer organization. Therefore, we considered repair TPT reduction only upstream in the supply chain.
- The same applies to the transportation TPT: reduction downstream is extremely expensive (helicopter flights) and therefore not a realistic option. Only TPT reductions upstream are feasible at reasonable costs.
- We have only two options for repair TPTs, namely either a normal or a fast repair. As shown in Section 6.2.3, this reduces the potential for cost savings.

7. Conclusions and directions for further research

In this paper, we developed a heuristic for the joint optimization of spare part inventories and TPTs of repair and transportation based on pricing of TPT reductions for multi-item, multi-echelon, multi-indenture spare part networks. Our heuristic is easy to apply and yields significant cost reductions compared to the standard VARI-METRIC method for spare part optimization where TPTs are fixed. We find that it is particularly profitable to reduce TPTs downstream in the supply chain. Repair TPT reduction of lower indenture items upstream in the supply chain is the least useful. In a case study at Thales Netherlands, we find a cost reduction of 5.6%, which is somewhat low compared to our theoretical experiments. This is due to the fact that TPT reductions downstream in the Thales network are very expensive because of the special business characteristics (an installed base of radar onboard of frigates).

Our approach is flexible and heavily relies upon the VARI-METRIC method for inventory optimization in multi-echelon, multi-indenture networks. As a consequence, we believe that known model extensions to VARI-METRIC can be included in our approach rather easily, thereby relaxing some model assumptions as mentioned in Section 3.1. For example, we can include the VARI-METRIC variants to deal with negative binomial demand (assumption 1), differences in item criticality (assumption 2), replenishment order quantities larger than 1 (assumption 8), and stochastic order-and-ship times (assumption 11) (Sherbrooke, 2004). Other model assumptions lead to considerably more complex models, in particular relaxing assumption 10 to include the use of lateral supply between stock points at the same level in
the multi-echelon structure. Even disregarding throughput time reductions, a complete approach for lateral supply in general multi-echelon, multi-indenture networks is still missing. Most models are limited to single or two-echelon networks with a single indenture level only (Paterson et al., 2011). This is a topic for additional research. We also expect that the impact of repair throughput time reductions may be less under lateral transshipments, because we have additional flexibility for fast supply via the lateral channels.

As other further research, we suggest developing a method for exact optimization of this model to provide a benchmark for the performance of our heuristic. The approach as applied by Basten et al. (under review b) for the joint optimization of the spare part provisioning and Level Of Repair Analysis (LORA) problem seems to be the most promising. However, we expect that an exact method requires more computation time, so that it will not be suitable to solve problem instances of practical size.

Acknowledgment

This research is part of the project on Proactive Service Logistics of advanced capital goods (ProSelO) and has been sponsored by the Dutch Institute for Advanced Logistics (Dinalog). We thank the Master student Maurice van Zwam for collecting case data at Thales Netherlands and performing the case study.

References