Repairable item and inventory control: a research proposal

Citation for published version (APA):

Document status and date:
Published: 01/01/1994

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.
Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
REPAIRABLE ITEM AND INVENTORY CONTROL

A RESEARCH PROPOSAL

Jos H.C.M. Verrijdt

Research Report TUE/BDK/LBS/94-13
February, 1994

Graduate School of Industrial Engineering and Management Science
Eindhoven University of Technology
P.O. Box 513, Paviljoen F16
NL-5600 MB Eindhoven
The Netherlands
Phone: +31.40.472247
Fax: +31.40.464596
E-mail: J.H.C.M.VERRIJDT@BDK.TUE.NL

This paper should not be quoted or referred to without the prior written permission of the author.

This research is sponsored by the Netherlands Organization for Scientific Research (NWO).
1. Introduction

In many companies the general goal is to achieve some pre-determined service performance (set by higher management) at minimum cost. When the external customer demand process is highly irregular, some kind of flexibility is needed to ensure the desired service performance at acceptable cost. Questions that often arise are: what sorts of flexibility can be distinguished (e.g. inventory, capacity) and how should it be applied to obtain maximum benefit at minimum cost?

The strong short-term fluctuations in the demand process, which is a main characteristic of the planning environment under consideration in this research report, can be identified in a number of situations. In general, one can identify two extreme types of companies that are faced with such highly varying demand processes:

1) Companies producing consumer goods in large quantities that are shipped all over the world to satisfy customer demand. These goods are often distributed through a number of consecutive warehouses (global, national, regional, local) before they arrive at the customer. The average demand is high and strong fluctuations in demand are caused by customers who order large quantities of goods at once.

2) After Sales Service organisations which have to replace failed components in complex technical systems (e.g. copiers, hardware systems, aircrafts) with spare parts. The failed parts are sometimes sent to a repair shop and if possible recovered. These organisations often also have a large distribution network consisting of consecutive stockpoints in order to be able to achieve a high service performance (i.e. minimal downtime of the technical systems). The demand for spare parts in these cases is characterised by a very low average demand with high variation.

In section 2 we briefly discuss the phenomenon of imbalance which is a main problem in companies of the first type. The rest of this report will focus on the second type of companies. In section 3 we give an overview of the developments in recent years in Spare Part Management. We also list some main characteristics of the spare part supply system. In section 4 a literature review is presented in which the most relevant Operations Research models for spare part control are discussed. Finally, in section 5 we give directions for the research to be carried out.

2. Imbalance

Companies facing a high average demand under strong short-term fluctuations, are sometimes confronted with inventory imbalance in their production/distribution network. This is caused by customers who order large quantities of goods at once (e.g. wholesale dealers). As a result, some stocking locations will be confronted with shortages whereas other locations are fully stocked. There are several ways of dealing with these undesirable imbalance situations. One could think of satisfying large portions of demand from upstream stockpoints (e.g. national instead of local stock). The demand process in the local stockpoints will therefore be smoothed. De Kok(1993) refers to this solution as large order overflow. Another way of dealing with large customer orders is splitting them into small portions that will be shipped to the customer in a number of consecutive periods. This order splitting procedure (cf. De Kok(1993)) will also smooth the demand process and reduce the variability. More research on these kind of multi-echelon networks can be found in De Kok(1990) and Verrijdt and De Kok (1993,1994).
3. Spare Part Management

In recent years an important shift towards customer service has taken place in the industries. Especially in highly competitive industries (e.g. automobiles, information systems, copiers), companies realize that offering extensive service to the customer can make a difference. The relation between supplier and buyer does often not end at the time of sale (Levitt (1983)). After sales service has become a competitive weapon. Long running service contracts with the customer force the supplier to react quickly whenever a product fails at a customer site. Defect parts must be swapped quickly with good parts, in order to minimize the customers down time. In order to respond adequately to customer calls, these companies often use an extensive spare part supply system. An efficient and effective service mechanism is vital for attracting new customers and keeping present customers at rebuy moments. Next to this tendency of focusing on customer service, economic developments force companies to reduce costs. Therefore, the main goal for many companies in the nineties is to improve the service performance and at the same time reduce the associated costs. To realize this goal, increasing attention is paid to the logistic structure of the spare part supply system. We now describe some general characteristics of such a service system.

3.1 Spare Part Categories

Spare parts can be categorized into two groups: consumables and repairables. Consumables are parts that can be used only once. After failure, these parts cannot be repaired and must be replaced by new parts. The failed parts are scrapped or recycled through disassembly. Repairables on the other hand can be repaired a number of times after they have failed. The failed part is sent to a repair facility where it is possibly recovered. Consumables are characterized by an one-way flow of parts through the supply system. Repairables are characterized by a closed loop: parts cycle through the supply system. In practice of course, this loop is not completely closed. Repairables can only be recovered a finite number of times and then have to be scrapped (thus leaving the loop). To replace these scrapped parts, new parts have to be purchased (thus entering the loop). It is especially these closed loop systems that have received much attention in research literature.

3.2 Network Design

A typical design structure of a spare part supply system is given in figure 1. Occurrence of a part failure in the installed base (i.e. collection of technical systems to be serviced) generates a spare part demand at the nearest service centre \( L_{ij} \). The failed part is returned to the repair shop for repair. After recovery the repaired part is shipped to a central warehouse \( C \) from where the regional depots \( R_k \) are supplied. The regional depots in their turn supply the local depots \( L_{ij} \) which are nearest to the customers. Sometimes a failed part cannot be repaired and must therefore be scrapped. As a result every now and then new parts have to be purchased from an external supplier. The return flow of failed parts mostly uses the same distribution network as the usable parts.
3.3 The Repair Shop

In figure 1 the repair shop is located centrally. All defective items are returned to one central facility for repair activities. It is, however, also possible that repair facilities exist at other stocking locations in the distribution network. Local stockpoints often have the ability to perform minor repair activities, whereas complex repair jobs are sent on to the central repair shop. In the literature one usually models this situation as follows: with probability $r$ a defective item is repaired at a local stockpoint and with probability $1-r$ a defective item is sent to the central repair shop.

Repair shops can have a hierarchical structure. This is due to the fact that the item to be repaired is hierarchically structured itself. An example, considering plane engines, will illustrate this (see figure 2). Plane engines consist of a number of replaceable modules which in their turn consist of a number of replaceable components (i.e. three layered structure). Whenever a plane engine fails, it is swapped with a serviceable engine in order to minimize the time a plane is grounded. The defective engine is sent to a repair shop and disassembled (layer one). Defective modules are swapped with serviceable modules (if available) and the engine is assembled and returned into serviceable state. The defective modules are also disassembled (layer two) and defective components are swapped with serviceable components. The modules are then returned via assembly into serviceable state. Finally, the defective components (layer three) are repaired and returned into serviceable state if possible. If repair is not possible, defective components are scrapped or recycled and external procurement is required to replace these components.

Determination of the repair throughput time of a defective engine in this situation can be very complex, since it depends on the availability of serviceable modules, which in their turn depend on availability of serviceable components.
3.4 Demand Process

The number of spare part types has increased enormously in recent years, due to an increased diversity of end products and shortened product life cycles. Tens to hundreds of thousands different spare parts in a service organization are no exception. The demand distribution for this wide range of spare parts is typically very skew (see figure 2). A small percentage of the parts, the fast movers, has a high demand (e.g. more than 10 per year), whereas most of the parts, the slow movers, face very low demand (as low as one demand per 10 years). These slow movers are usually very expensive and therefore represent a major factor in inventory investment in spare parts.

The demand for spare parts is generated by machine failures at customer sites. It is very difficult, especially for slow movers, to forecast the demand, because the failure of machines is a stochastic process with high uncertainty. The strong short term fluctuation in demand is typical for a spare part environment.
3.5 Service Measures

There are essentially two different ways of measuring service performance in a spare part environment. We can distinguish between a time-weighted performance measure and a quantity-weighted performance measure. A time-weighted service measure is usually expressed in terms of a maximum response time:

$$\Pr(\text{response time} < t \mid \text{demand occurs}) > \beta$$

where $\beta$ and $t$ are given input parameters. A number-of-incidents measure is usually expressed in terms of fill rate: the fraction of customer demand that can be satisfied from stock on hand. Both measures are used in practice. In spare part environments, however, a time-weighted service measure (e.g. maximum response time) is to be preferred for the following reason (see Muckstadt and Thomas, 1980).

Consider a two-echelon distribution structure consisting of a central warehouse that supplies a number of regional stocking locations, situated nearby customers. If a particular spare part item is not stocked at these regional stocking locations (e.g. the part is very expensive and the demand rate is very low), the central warehouse has no incentive to stock that particular part if a fill rate measure is used. A spare part demand at a regional facility is always backordered (i.e. fill rate is zero), independent of the central inventory. However, when a response time measure is used, the central warehouse does have an incentive to stock that particular part, since it influences the response time for a customer call.

Cohen and Lee (1990) mention five practical service measures. Measures (1) to (4) are quantity-weighted measures whereas measure (5) is time-weighted:

1) **Part Unit Fill Rate:** the fraction of demand delivered from inventory on hand over some review period.

2) **Part Dollar Fill Rate:** identical to 1) except that items are weighed in money value instead of units.

3) **Order Fill Rate:** fraction of internal replenishment orders that can be completely filled from inventory on hand.

4) **Repair Order Completion Rate:** fraction of repair jobs at customers sites that are not delayed by part shortages.

5) **Customer Delay Time:** incurred delay between the identification of a service need (customer call) and satisfying that need. Many companies for example guarantee service within 4, 8 or 24 hours.
Important to note is that in practice the service performance is measured only at the various stocking locations. The service performance experienced by the customer is seldom known.

3.6 Control Aspects

The flow of parts through the system can be controlled in several ways. The inventory policy applied at a stocking location is usually controlled locally. Stocking locations often represent independent organizational units with their own financial responsibility (e.g. National Sales Organizations). This makes it difficult to apply integral multi-echelon approaches, although these approaches can achieve considerable savings in inventory investment (see Muckstadt and Thomas(1980)). The inventory policy is usually 141 (one-for-one) replenishment for slow movers and \((s,Q)\) replenishment for fast movers.

Other important control aspects are:

**Sourcing:** Who can order where? Which supplier-buyer relations are possible for normal replenishment orders as well as for emergency orders? A special case of sourcing is pooling between end depots. When one depot faces a shortage and a nearby situated pooling-depot has a surplus, it could be efficient to reallocate inventory.

**Transportation modes:** Choosing between different forms of transportation is a cost-service trade-off. Instant delivery (i.e. high service) for example can be assured by using fast but expensive transportation services such as planes.

**Prioritization:** Not all customers are equal. Depending on service contracts, some customers will enjoy priority treatment. This differentiation of customers can also be an important control parameter.

3.7 Product Life Cycle

A well known problem in the field of spare part management is the determination of initial stocks. When a new product is introduced into the market, spare part inventories are required to support service logistics. However, the determination of these spare part inventories is very difficult, since no historical data on failure rates are available. Information concerning demand for spare parts only becomes available in the course of time. Consequently, initial inventories are usually estimated with the help of technicians, who have experience with similar parts. These initial estimates, however, can deviate substantially from the true demand.

A similar problem arises when the manufacturing of a product is terminated. A spare part manager is often confronted with a "last call" opportunity to order spare parts. Ordering of spare parts at a later moment in time is usually very expensive or impossible, since it requires the rebuilding of a former production line. When faced with this "last call" problem, a manager has to estimate the demand for spare parts for a period of time guaranteed in the various service contracts with customers. Ordering too much contains the risk of obsolescence, whereas ordering too little contains the risk of an expensive order at a future point in time.

4. Literature Review

When looking at multi-echelon base-depot supply systems for repairable items, the METRIC-model (Multi-Echelon Technique for Recoverable Item Control, Sherbrooke 1968) is widely considered to be the first model that captures the most important features of the problem of determining inventory levels for spare parts in a multi-echelon environment. It was successfully implemented at the US Air Force. METRIC is a mathematical model that consists of a central depot supplying a number of bases with various types of recoverable parts. All parts are assumed
to be repairables and therefore no external procurement is allowed. The demand for spare parts is generated at the bases and is assumed to be compound Poisson. A defective part that is returned at a base is immediately replaced by a spare part from stock on hand at the base (or backordered when no stock is available). The defective part is repaired either at the base or at the depot. When the part is sent to the depot for repair, an immediate resupply order is generated for that part at the depot (i.e. 141-replenishment). When the depot has serviceable stock on hand, a spare part is shipped to the base. Otherwise, a spare part is backordered and will be shipped as soon as it becomes available from the repair process. METRIC determines spare part inventory levels for all parts at all stocking locations (depot and bases) that minimize the sum of the expected backorders of all parts at all bases at a random point in time, subject to a investment constraint.

An important assumption that is made in METRIC is that repair times for all parts are independent (i.e. there is no waiting or batching of defective parts). This infinite capacity assumption enables Sherbrooke to apply Palm’s Theorem (Palm, 1938) which states that if failures are generated by a stationary Poisson process and repair times are independent, identically distributed random variables, then the steady-state number of parts undergoing repair at any given time is also Poisson with a mean equal to the product of the failure rate and the mean repair time. Feeney and Sherbrooke (1966) showed that Palm’s Theorem is also applicable for compound Poisson failure processes. The importance of this extension of Palm’s theorem lies in the fact that with compound Poisson distributions one can obtain variance-to-mean ratios greater than one, whereas the simple Poisson process has a variance-to-mean ratio exactly equal to one.

Since the development of METRIC a lot of research has been done on multi-echelon repairable-item inventory models. In reviewing the extensive literature (see Nahmias (1981), Mabini and Gelders (1990) and Cho and Parlar (1991)), we distinguish two categories (cf. Cho and Parlar, 1991). First, we consider METRIC-based models, that are characterized by (compound) Poisson failure processes and the infinite capacity assumption as stated above. These models disregard repair shop restrictions such as a finite number of repair men, batching policies for defective parts, and priority scheduling in the repair process. The primary focus is on determining optimal stocking levels subject to some cost or service criteria. An excellent description of the METRIC-approach is given in Sherbrooke (1992). Secondly, we consider the non-METRIC models, which do take account of the repair shop restrictions as mentioned earlier. A lot of research on these models is based on queuing theory. The main difference with the METRIC-based models is that repair capacity is considered a control variable. Finally, we consider a series of papers by Cohen et al. that do not take account of the possibility of repair (only consumables are considered) but that have a very practical significance.

4.1 METRIC-based models

The first essential extension to the METRIC model is the multi-indenture relationship between end items and their comprising modules or components. METRIC does not take account of the hierarchical structure of these end items. In the Air Force, where METRIC was implemented, this means that no distinction is made between spare engines (end items) and the comprising modules. In determining optimal stock levels for spare parts, METRIC minimizes the expected backorders of all items. In practice, however, only shortages of end items (i.e. engines) affect the downtime of technical systems (i.e. planes) immediately. A shortage of modules has no direct impact on the downtime of a technical system.

Sherbrooke (1971) was the first to recognize this multi-indenture relationship and he developed an expression for the expected backorders of end items. He assumes that an end item consist of a number of replaceable modules. Failure of an end item is caused by one module i (with probability p_i) or some other cause (with probability p_0). Failed modules are repaired
without delay, i.e. spare parts that are needed to repair the modules are immediately available. The model is for a single base and is evaluative in nature.

Muckstadt(1973) extended METRIC to a multi-indenture model that he called MOD-METRIC. He derives an expression for the mean base repair time, consisting of a mean repair time plus a delay due to the unavailability of modules. MOD-METRIC is an optimization model (like METRIC) that determines stock levels for all parts at all locations that minimize the expected base backorders of the end items, subject to an investment constraint. The model was implemented by the US Air Force for the F-15 weapon system.

The METRIC model is completely conservative. Failed parts can be repaired an infinite number of times. Simon(1971) extends the METRIC model to allow for positive condemnation rates: parts can be repaired (at the bases or at the depot) but they can also be condemned (repair is not longer possible). The depot applies an \((s,S)\) strategy: when the inventory drops below \(s\), an external procurement order is issued to increase the inventory level up to \(S\). Because the bases apply 141-ordering policies, the depot effectively applies an \((s,Q)\) ordering policy, where \(Q=S-s\). Simon derives exact expressions for expected backorders and stock on hand and in repair at each stocking location. Restrictions to the model are that replenishment times are assumed deterministic and that demands at the bases follow simple Poisson processes (i.e. variance to mean ratio equals one).

In both METRIC and MOD-METRIC, demands from bases placed upon the depot are filled using a FCFS-rule. Miller(1974) models another rule: an item that has completed depot repair will be shipped to that base \(j\) whose marginal decrease in expected backorders will be the greatest at \(T_j\) days into the future (where \(T_j\) represents the constant transportation time from depot to base \(j\)). Miller, however, makes the following restrictive assumptions: demands are simple Poisson processes and repair times are independent exponential random variables.

METRIC assumes that the number of outstanding orders at the depot (i.e. pipeline stock) of each base is Poisson distributed. However, it can be shown that the variance of the pipeline stock distribution exceeds the mean (except when all base stock levels are zero). The error that is made when assuming a Poisson distribution can be significant. The models by Slay(VARI-METRIC, 1984) and Graves(1985) correct this deficiency by using two-moment approximations. Graves considers a repairable item base/depot supply system where repair is only possible at depot level. Demand at the bases is compound Poisson and shipment times from depot to bases are deterministic. Graves presents an exact model for the steady state distribution of the outstanding orders at the bases and he presents an approximation model in which the first two moments of this distribution are fitted to a negative binomial distribution. This approximation model performs better than METRIC in a set of test problems.

Sherbrooke(1986) states that multi-indenture models such as MOD-METRIC underestimate the delay in repair of end items (due to shortages of components), by assuming a Poisson distribution for the number of end-items in resupply. These models also underestimate the incurred delay in resupply as a result of backorders at the supply center (as shown by Slay(1984) and Graves(1985)). Sherbrooke applies the two-moment approximation method (VARI-METRIC) to the multi-indenture multi-echelon system and he shows improved performance compared to MOD-METRIC. In fact, the performance is very close to the "true" simulation results.

Kaplan and Orr(1985) present a model called OATMEAL (Optimum Allocation of Test Equipment/Manpower Evaluated Against Logistics). Most multi-echelon repairable item models use the maintenance policy (what to repair where?) as input and determine optimal stockage policies. OATMEAL, however, determines simultaneously optimal maintenance as well as stockage policies for a weapon system. The objective is to minimize total inventory investment subject to a target level for the operational availability of the system. The model uses Mixed-integer programming combined with a Lagrangian approach.

Dada(1992) models a single-item multi-echelon spare part system with priority shipments.
Each of N bases stocks exactly one spare item. Demand at the bases is generated by independent Poisson processes. The bases reorder at a central warehouse that stocks m items. The central warehouse reorders at an outside vendor. If demand occurs when a base is out of stock, a number of priority shipment options is available to fill this demand. An aggregate model and a decomposition method to approximate steady state distributions are developed.

4.2 Non-METRIC models

All research discussed so far uses the infinite capacity (or ample server) assumption: repair times are independent identically distributed stochastic variables and failed parts are immediately taken into repair upon arrival at the repair shop (no queuing). In practice, however, repair capacity is limited and as a result repair times are correlated and waiting times can occur. A trade-off has to be made between the number of repair men (capacity) and the level of spare parts (inventory). We now discuss some relevant work that assumes limited repair shop capacity.

Mirasol (1964) considers a single-echelon single-repair shop with finite repair capacity and an infinite source generating demands for spare parts. He makes a trade-off between number of spares and the number of repair channels using a system unavailability criterion.

Gross (1982) considers a single-echelon, single-item model with a limited number of repair channels for failed items. He compares the M|M|∞ system (i.e. the METRIC model with infinite repair capacity) with a M|M|c system (i.e. only c repair channels are available). Numerical results are presented that show the error that is made when assuming ample service when it actually is not. It is shown that the error that is made can be sizable for high values of the demand rate, a low number of repair channels, or high target values for the service performance.

Scudder and Hausman (1982) use a simulation model for a repair shop with limited capacity (i.e. dependent repair times) to evaluate the performance of several scheduling priority rules for multi-indentured repairable items. They compare their simulation model with the MOD-METRIC model which assumes infinite capacity (i.e. independent repair times). Their results show that models assuming infinite repair capacity when it actually is not, perform quite good.

Hausman and Scudder (1982) evaluate the performance of different priority scheduling rules in a finite capacity repair shop, supporting a repairable inventory system with a hierarchical product structure. The product considered is a jet engine, consisting of different modules, which in their turn consist of different components. A failure of an engine is caused by exactly one module and a failure of a module is caused by exactly one component. The authors assume a constant repair shop capacity and a constant inventory investment in spares. They simulate a repair shop with ten work centers to evaluate the performance of three classes of priority scheduling rules for the repair of components. The performance criterion used is the expected delay days for engines. The three classes of priority rules are: static rules (independent of job status and inventory status), dynamic rules that depend on job status and dynamic rules that depend on job status as well as inventory status. This last class of rules proves to be superior to the other two.

Scudder (1984) extends the former model to the multiple-failure case. In practice, dependent failures occur when failure of one module (component) may trigger the failure of another module (component). A group of modules (components) which are likely to fail simultaneously are referred to as clusters. The results show that priority rules that perform well for the single-failure case also perform well for the multiple-failure case. More complex rules, incorporating clustering characteristics, do not appear to provide any significant improvement.

Gross, Miller and Soland (1983) extend the model of Mirasol (1964) in the following way: demands are not generated by an infinite source but by a finite number of operating machines, and a 2-echelon environment is considered, consisting of one depot and one base. Repair activities can take place at both echelons, whereas stocking of spare parts is only possible at base level.
Repair and failure rates are assumed to be exponentially distributed. The model minimizes total expected costs subject to a pre-specified system availability constraint.

In Gross and Miller (1984) the former model is treated in a time-varying environment: repair and failure rates vary over time. They also allow for more bases to be modelled and spare part stocking at both echelons. The model is compared with Dyna-METRIC (Hillestad, 1981) in which infinite capacity and an infinite source of demand is assumed.

Gross, Kioussis and Miller (1987a,b) use a randomization technique and a truncated state space approach to handle successfully large Markovian finite source, finite repair capacity models. Their work is an extension of Gross and Miller (1984).

Albright and Soni (1988) studied another two-echelon repairable-item inventory system with a central depot and N repair bases with exponentially distributed failure and repair rates. Spares are only stocked at the bases. The return policy for repaired items prescribes that an item has to be returned to that base which has the highest percentage of its units in the depot.

Ebeling (1991) analyzes a single-echelon multi-item repairable inventory system. His model consists of L operating systems, each consisting of M different items. The objective is to determine optimal values $S_i$ (spare stock level) and $K_i$ (number of dedicated repair channels) for each item $i$ that maximize the system availability, subject to a budget constraint. System availability is defined as the product of the ready rates of the different items. In calculating the ready rates, Ebeling uses the steady state probabilities of a $M | M | K_i$ queue. The optimization method consists of two steps. First, for each item $i$ and for each feasible budget, the values $S_i$ and $K_i$ are determined. Secondly, using dynamic programming, the total budget is allocated among the M items.

Daryanani and Miller (1992) analyze a single-item multi-echelon repairable inventory system with one central repair facility and several bases. The repair facility has one repair channel with exponentially distributed repair times. The failed items arrive at the repair shop according to independent Poisson processes. The main objective of the paper is to compare a dynamic return policy for repaired items (i.e. the base with the highest number of replenishment orders has priority) with the standard FCFS-policy. The technique used to calculate the steady state probabilities is called taboo structure. The performance measure used is base availability. The dynamic return policy is superior to the FCFS-policy and can affect the performance substantially.

Finally, Buyukkurt and Parlar (1993) conducted a simulation study to evaluate two static and one dynamic allocation policies (i.e. depot-to-base return policy for repaired items): (1) return an item to the base which has the longest outstanding backorder, (2) return an item to its original base, and (3) return an item to the base which needs it most. The dynamic policy proved to be superior to the two static policies under five different optimality criteria.

4.3 Cohen models

We now focus on a series of papers by Cohen et al. These authors address the problem of determining optimal stocking levels for spare parts in multi-echelon systems, applying a periodic review inventory policy. The goal is to minimize total costs subject to some service level constraint. These models cannot be classified as METRIC or non-METRIC models, since all spare parts are assumed to be consumables. Failed parts are replaced by serviceable parts that are stocked at various locations. The failed parts are scrapped instead of being returned to a (central) repair facility. Their work, however, is very interesting since they take account of a number of real-life aspects, such as emergency transhipments, demand priorities, and pooling mechanisms in a multi-echelon environment. The model they developed was implemented at IBM's US after sales service logistics system with great success (see Cohen et al. 1990).

The first paper (Cohen et al. 1986) considers a single-item multi-echelon divergent distribution network for spare parts. Demand for spare parts originates at the lowest level (echelon) of the tree structure. Excess demand that cannot be met from stock on hand is passed
on to the supplying stockpoint at the next higher echelon. This excess demand is therefore considered as lost to the lower stockpoint. The model also allows for the possibility of pooling. Stocking locations at the same echelon are divided into pooling groups. Before sending excess demand in a particular stockpoint $i$ to a higher echelon stocking location, it is first checked whether neighbouring stocking locations, belonging to the pooling group of location $i$, have excess inventory after their demand is satisfied. If so, this excess inventory is used to meet the excess demand at location $i$. If not, or the excess inventory of the pooling group is insufficient to meet all excess demand, the remaining excess demand is passed on to a higher echelon stocking location. This procedure is repeated at each echelon.

The model also allows for stock recycling at the lowest echelon. Parts that are demanded are returned to stock with a given probability, due to the fact that they were only needed for diagnostic use. The objective is to determine optimal stocking levels for each location, that minimize total expected costs (i.e. emergency shipments, normal replenishments, and inventory holding) per period, subject to a response time constraint (e.g. 95% of the demand is fulfilled within 4 hours). Restrictive assumptions in the model are: deterministic lead times (1), independent failure processes at the lowest echelon (2), emergency shipment costs only depend on the supplying location (3), and most important no outstanding backorders exist at the beginning of a review period (4). Assumption (4) is motivated by the fact that the demand rates are extremely low, and therefore normal replenishment times are much smaller then average demand interarrival times.

In Cohen et al. (1988) a single-echelon single-item $(s,S)$ inventory system under periodic review is described. The demand imposed at the stocking location consists of two priority classes: emergency demand (with high priority) and normal replenishment demand (with low priority). Excess demand that can not be met from stock on hand is satisfied through an emergency procedure and is considered lost to the stocking location. The replenishment lead time for the stocking location is fixed and the demand distribution is known. A heuristic is developed to determine the values $(s,S)$ that minimize total expected costs subject to a fill rate constraint (i.e. fraction of demand filled from stock on hand $\geq \beta$). Total costs consist of ordering costs, holding costs, transportation costs, and shortage costs (determined by the cost of the emergency procedure). In a later paper (Cohen et al. 1992) this model is extended to a multi-item environment where the service constraint is defined at product level, not at item level.

A single-echelon multi-item periodic review inventory system is considered in Cohen et al. (1989). A product consisting of $n$ parts (field replaceable units) is serviced by a stocking facility that stocks all $n$ parts. The objective is to determine order-up-to-levels $S_i$ for all parts $i$, $i=1...n$ at the stocking facility, such that expected total costs are minimized subject to a service constraint. Costs consist of ordering costs, holding costs, transportation costs, and shortage costs. Excess demand is satisfied through an emergency procedure (which determines the shortage costs) and is considered lost to the stocking facility. The service constraint is defined at product level, not at part level. A chance constraint (fraction of the time that all demanded parts can be delivered from stock on hand) and a parts availability constraint (weighed fraction of parts that can be delivered from stock on hand) are considered. The following assumptions are made: the facility is fully stocked at the beginning of a review period (i.e. demand rates are very low), no priority classes of demand exist, and no distinction is made between "real" usage and diagnostic usage (in which case the part is returned to stock). The paper also extends the above model to a two-priority demand classification and commonality between parts (i.e. a single part can be used in different end products).

5. Directions for further research

When looking at spare part systems as described in section 3, there are several ways of...
influencing the logistics of spare part flows. In this paper we distinguish two main types of flexibility that can be used to affect the system performance. In section 5.1 we discuss these types of flexibility. In sections 5.2 and 5.3 we consider different circumstances in which these types of flexibility can be applied.

5.1 Types of flexibility

In this research paper we will try to give an overview of ways to increase the performance of a spare part system by creating flexibility. In general, we distinguish two types of flexibility that can be applied to increase system performance:

1) Capacity Flexibility: The function of a repair shop in a spare part environment is to return defect parts (or modules, components etc.) into serviceable state. In order to be able to develop a mathematical model for such situations, one often makes simplifying assumptions. The METRIC-like models, for example, assume that an infinite number of repair channels is available at the repair centre, implying that no waiting queues of defect parts arise and that repair times are mutually independent. Another simplifying assumption one often sees in the literature is the FCFS-priority assumption: repair jobs are dealt with in the sequence in which they arrive. Simplifying assumptions, such as the ones stated above, restrict the flexibility of the repair shop. It is possible to create flexibility in the repair shop by relaxing these restrictive assumptions. We now list a number of capacity flexibility options:

* time varying repair capacity: e.g. increase the capacity (by means of overtime or extra hired help) in busy times (see De Haas(1994))
* priority scheduling of repair jobs: the sequence in which jobs are repaired can depend on some priority scheme (e.g. emergency jobs always have priority over other jobs)
* batching: instead of repairing jobs on a 141-basis, it could be sensible to work in batches

2) Inventory Flexibility: Another way of creating flexibility in a spare part system is inventory control. Most models in literature concentrate on determining optimal stock levels for spare part types in all stocking locations, that satisfy some target performance subject to a budget constraint. The determination of these optimal stock levels is usually done only once and for a very long period. Although these stock levels are fixed for all locations, one still can obtain flexibility in the following ways:

* pooling: instead of supplying end stockpoints (i.e. stockpoints which are located the nearest to the customers) centrally from a regional warehouse, it is also possible to tranship parts laterally from other end stockpoints located nearby. Especially in emergency situations, this solution is often very practical.
* allocation policies: whenever a part is recovered and leaves the repair shop, it is possible to apply a number of criteria that determine to which location the repaired part has to be sent. The same holds for central stocking locations that supply a number of downstream stockpoints.

As we have seen, the system performance can be influenced in a number of ways. We now distinguish between two situations: normal replenishment and emergency replenishment. Normal replenishments occur whenever the inventories in one or more stocking locations in the network drop below the stocknorms for the particular locations. The supplier-customer relationships between stocking locations for these kind of replenishments are usually fixed. Every stocking
location orders at a pre-determined location at a higher echelon. Emergency transhipments occur whenever one or more stocking locations are faced with a backorder situation. Especially when a backorder occurs at a stocking location which directly supplies external customers, adequate emergency actions are required to minimize the customers down time and to maximize service performance. In the next two subsections we give an overview of the possible control parameters to increase flexibility for these two kind of replenishments. The type of flexibility that can be applied for normal replenishments exist of a mixture of both capacity flexibility and inventory flexibility. For emergency replenishments capacity flexibility is usually not an option, since immediate action is required. Waiting for a failed item to be repaired is not allowed, because service contracts require a short response time when a machine at a customers site is down.

5.2 Normal replenishment

When discussing the flexibility control parameters for normal replenishments, we follow figure 4.

The first opportunity to increase flexibility is the order release function for the repair shop. The following question arises: what item type should be taken into repair when a repair channel becomes available? This priority scheduling problem is of course only relevant for the multi-item situation.

Several priority scheduling rules can be applied: First Come First Served, Longest Queue etc. A priority rule that takes account of the total system echelon inventory is the following: take item \( i \) in repair that has the highest percentage \( p_i \) of its initial system stock \( Y_i \) in the repair shop. An example will illustrate this strategy. Suppose we have 3 item types. The initial system stock is chosen as follows: \( Y_1 = 10, Y_2 = 7, Y_3 = 5 \). The system stock consists of the sum of all initial inventories in all stocking locations in the distribution network. At a given moment a repair channel becomes available in the repair shop. Suppose that the repair shop inventory \( X_i \) (i.e. items of type \( i \) waiting for or in repair) are as follows: \( X_1 = 3, X_2 = 2, X_3 = 4 \). The percentages \( p_i \) are then computed as follows: \( p_1 = 0.30, p_2 = 0.29, p_3 = 0.80 \). This implies that item type 3 should be taken into repair next. Not that it is possible that \( p_i \) can become larger than one. This is the case when \( X_i \) is larger than \( Y_i \), implying that one or more customers (machines) in the field are down and waiting for a serviceable item of type \( i \).

The second opportunity to influence the system performance is the flexibility of the repair shop capacity. By increasing the repair capacity when the number of failed items, waiting for repair, becomes very large, the system performance can be influenced. Increasing capacity when needed for a certain time period can be achieved in a number of ways. One could think of hiring temporary employees, overtime work of repair men, or subcontracting of repair work. The policy that dictates when and how much to increase capacity, should reflect in some way the current
repair work content of the repair shop. De Haas(1994) considers a single-item $M|M|c$ queue where capacity in increased when the queue length exceeds a certain threshold. This critical queue length is determined by a target level $r$, which represents the fraction of overtime increased capacity is desired.

The third possibility to influence system performance is the return policy for repaired items. Upon completion of repair of a failed item, one can send the repaired item to several destinations. The choice to which location the item should be send influences the system performance. In most literature one assumes a FCFS-policy: the location with the longest outstanding backorder receives the item. Another interesting policy to consider is a policy that takes account of the echelon inventories per item type in the different branches of the distribution network. The example depicted in figure 5 will illustrate this policy.

Repaired items that leave the repair shop can be distributed to 3 continental warehouses in Europe (E), North-America (NA), and Asia (A). Consider the single-item case. Extension to the multi-item case is however straightforward. The initial spare part inventory is set as follows: $Y_E = 20, Y_{NA} = 20, Y_A = 15$. These inventories include all spare parts located at the continental warehouses and at the regional warehouses further downstream in the distribution network (Eindhoven, Marseille, Dallas etc.). At a given moment a repaired item leaves the repair shop. The echelon inventories $Z_i$ of the 3 branches at that moment are as follows: $Z_E = 15, Z_{NA} = 8, Z_A = 10$. The allocation policy dictates that the repaired item should be sent to the continental warehouse which has the lowest percentage $q_i$ of its original system stock currently available. In this case: $q_E = 0.75, q_{NA} = 0.40, q_A = 0.67$. So the repaired item will be sent to the continental warehouse in North-America.

![figure 5: example distribution network](image)

This allocation policy can be applied in a similar way at the continental warehouses. At the time of allocation, determine for every continental warehouse the regional warehouse which has the lowest percentage of its original stock currently available. The item will be allocated to that particular warehouse.

Not all items will be distributed throughout the distribution network. It is of course possible that some of the item types will only be stocked centrally at the repair center or at the continental warehouses. This decision depends on the demand and cost characteristics of the various item types.
5.3 Emergency replenishment

When a backorder occurs at a stocking facility that supplies external customers, this means that one or more customers are down. At such moments, emergency actions are required to minimize the downtime of the customer. Many service contracts include the condition that, at a moment of failure, service will be procured within a given amount of time. Normal replenishment is not an option in these situations, since it usually requires too much time. The backordered item must be supplied immediately from an emergency source. The possible emergency replenishment actions are illustrated with the help of figure 5. Suppose a customer backorder occurs at the warehouse in Dallas. Two types of emergency supply sources are considered:

1) Higher echelon source: The backordered item can be shipped from the continental warehouse of North-America at the next higher echelon. If the item is not available at the continental warehouse, there is also the possibility to ship the item from the central repair shop inventory. These kind of emergency transhipments are characterized by the fact that they use the same routings as the normal replenishment shipments.

2) Pooling source: Instead of using the normal replenishment transportation channels for emergency transhipments, lateral transhipments or pooling might also be worth considering. The Dallas warehouse can check at the Chicago warehouse for excess inventory. In this way, the inventories in the distribution network are balanced. The continental warehouses can also apply this possibility. For example, the continental warehouse in North-America can check for excess inventory at the continental warehouses in Europe and Asia.

Choosing between all these possible emergency replenishment options depends on a trade-off between response time requirements and transhipment costs.

References


Miller, B.L. (1974), "Dispatching from depot repair in a recoverable item inventory system: on the optimality of a heuristic rule", Management Science 21, 316-325.


Nahmias, S. (1981), "Managing repairable item inventory systems: a review", in: L.B. Schwarz


