

Design and control of service part distribution systems

Citation for published version (APA):

Verrijdt, J. H. C. M. (1997). *Design and control of service part distribution systems*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences]. Technische Universiteit Eindhoven. <https://doi.org/10.6100/IR475533>

DOI:

[10.6100/IR475533](https://doi.org/10.6100/IR475533)

Document status and date:

Published: 01/01/1997

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:

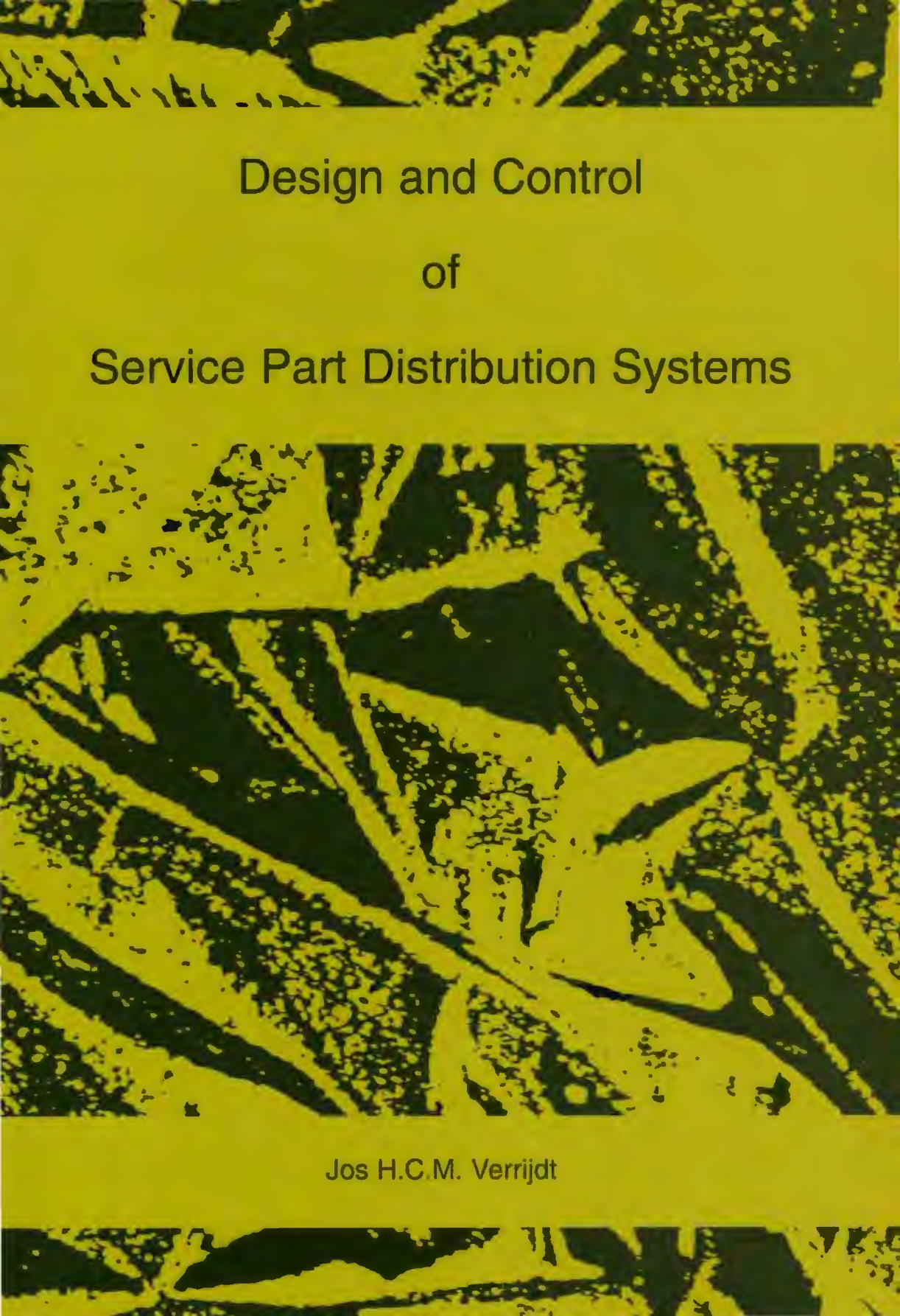
www.tue.nl/taverne

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.



Design and Control
of
Service Part Distribution Systems

Jos H.C.M. Verrijdt

DESIGN AND CONTROL
OF SERVICE PART
DISTRIBUTION SYSTEMS

**DESIGN AND CONTROL
OF SERVICE PART
DISTRIBUTION SYSTEMS**

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van
de Rector Magnificus, prof.dr. M. Rem,
voor een commissie aangewezen door het College
van Dekanen in het openbaar te verdedigen op
donderdag 30 januari 1997 om 16.00 uur

door

Jozef Hubertus Catharina Maria Verrijdt

geboren te Well

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. A.G. de Kok

en

prof.dr. J.A.M. Theeuwes

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Verrijdt, Jozef Hubertus Catharina Maria

Design and control of service part distribution systems /

Jozef Hubertus Catharina Maria Verrijdt.

Thesis Technische Universiteit Eindhoven. - With ref. -

With summary in Dutch.

ISBN 90-386-0325-8

NUGI 689

Subject headings: inventory control / service part
management / distribution control



Druk: Ponsen & Looijen, Wageningen

© 1997, J.H.C.M. Verrijdt, Eindhoven

PREFACE

In December 1992 I started working on my PhD research. In some way, it was an adventure of which the outcome, and the path that led to it, was surrounded with uncertainties and challenges. Looking back now, after four years of working at the department of Operations Planning and Control of the faculty of Technology Management at the Eindhoven University of Technology, I feel that these four years have contributed enormously to my personal development. Especially, I enjoyed the variety of activities I was involved in. The combination of scientific research, education (e.g. supervising master projects), and contacts with companies was very appealing to me. I also enjoyed the activities that were not directly related to my research but that were very interesting and stimulating. Being a member of Studium Generale (organizing weekly lectures on current affairs), the LAIOOB (Dutch network of PhD students in the field of industrial engineering and management science), and the editorial board of SCOPE (magazine of the faculty of Technology Management) was very rewarding.

This book is the product of four years of scientific research. Many people and organizations have supported me in writing this thesis. The Netherlands Organization for Scientific Research (NWO) provided the financial and material resources for conducting this research. Ton de Kok, my advisor during the last five years in which I worked on my master thesis and PhD thesis, provided invaluable support. He generated a never ending stream of ideas, suggestions, and comments that contributed much to the contents of this thesis. I would also like to thank Rommert Dekker, Jacques Theeuwes, and Henk Zijm for their comments and suggestions in the final stage of the writing process. A special word of thanks goes to two persons with whom I worked together intensively in recent years: Ivo Adan (faculty of Mathematics and Computing Science, Eindhoven) and Patrik Alfredsson (Royal Institute of Technology, Stockholm). The results of these cooperations constitute a significant part of this book. A pleasant working environment is essential in any job. The (former) members of the department of Operations Planning and Control provided such an environment for which I am very grateful. I would like to thank in particular Jan Fransoo, my office mate, who endured my presence during the periods he was not abroad. Furthermore, I thank the members of the Parts Business Forum for providing a practical frame of reference with respect to the ideas and models presented in this thesis. Last, but certainly not least, I thank my girlfriend, my family, and all my other friends for their support and interest in recent years.

Jos Verrijdt

Eindhoven, November 1996

TABLE OF CONTENTS

| | | |
|------------------|--|-----------|
| Chapter 1 | Introduction | 1 |
| 1.1 | Motivation for this study | 1 |
| 1.2 | Service Part Supply System: an example | 3 |
| 1.3 | Characteristics of service part logistics | 4 |
| 1.4 | Objective of this thesis | 15 |
| 1.5 | Outline of this thesis | 17 |
| Chapter 2 | Literature review | 19 |
| 2.1 | Introduction | 19 |
| 2.2 | The METRIC model | 19 |
| 2.3 | Extensions to the METRIC model | 21 |
| 2.4 | Cohen models | 24 |
| 2.5 | Repair models | 25 |
| 2.6 | Positioning and discussion | 28 |
| Chapter 3 | Service Part Supply System: a framework for control | 31 |
| 3.1 | Introduction | 31 |
| 3.2 | The Service Part Supply System | 31 |
| 3.3 | Economic trade-offs | 35 |
| 3.4 | Embedding the SPSS framework | 39 |
| 3.5 | Case studies | 42 |
| 3.6 | Conclusions | 54 |
| Chapter 4 | The Emergency Repair Model | 55 |
| 4.1 | Introduction | 55 |
| 4.2 | Related literature | 57 |
| 4.3 | Model description | 59 |
| 4.4 | Model analysis | 63 |
| 4.5 | Numerical evaluation | 71 |
| 4.6 | Comparison with the Muckstadt-Thomas model | 85 |
| 4.7 | Conclusions | 87 |

| | | |
|-------------------|---|------------|
| Chapter 5 | The Emergency Supply Model | 89 |
| 5.1 | Introduction | 89 |
| 5.2 | Related literature | 90 |
| 5.3 | Model description | 92 |
| 5.4 | Model analysis | 96 |
| 5.5 | Model validation | 101 |
| 5.6 | Economic evaluation | 105 |
| 5.7 | Pooling structures | 107 |
| 5.8 | A practical example | 113 |
| 5.9 | Conclusions | 120 |
| Chapter 6 | Policy evaluation | 123 |
| 6.1 | Introduction | 123 |
| 6.2 | The distribution network configuration | 123 |
| 6.3 | Parameter settings | 125 |
| 6.4 | Policy descriptions | 127 |
| 6.5 | Design issues | 129 |
| 6.6 | Numerical results | 129 |
| 6.7 | Conclusions | 141 |
| Chapter 7 | Conclusions and recommendations for further research | 143 |
| 7.1 | Introduction | 143 |
| 7.2 | Main conclusions | 144 |
| 7.3 | Topics for further research | 148 |
| References | | 151 |
| Appendices | | 157 |
| Appendix A | | 157 |
| Appendix B | | 162 |
| Appendix C | | 164 |
| Appendix D | | 166 |
| Appendix E | | 169 |
| Appendix F | | 171 |
| Appendix G | | 173 |

| | |
|--|------------|
| Summary | 175 |
| Samenvatting (Summary in Dutch) | 179 |
| Curriculum vitae | 183 |

Chapter 1

Introduction

1.1 Motivation for this study

During this century an important shift has taken place in industry in general. Until the sixties the market for many products could be characterized as a suppliers market. The demand for products was so high that companies had little problem selling their products to customers. Many manufacturing environments were characterized by large production batches and a relatively small diversity of products. This situation, however, has changed dramatically since then. Due to market saturation and increasing global competition in many industries, the suppliers market changed to a buyers market. Customers became more demanding and the variety of products increased enormously. Due to this changed situation companies faced higher and more aggressive competition. In order to attract new customers and to keep current customers, more attention is given to the After Sales activities of companies. For example, in 1988 the After Sales service market for electronics companies was estimated to be over \$20 billion and provided up to 15-25% of the total revenues of these companies. Moreover, this service market was expected to grow to nearly \$50 billion in the nineties and provide up to 30% of the total revenues (Hull and Cox, 1994). In a field study 'Service Operation Strategies of the 90s' (Coopers & Lybrand, 1991) it is also recognized that "*...It is a service of excellence which differentiates products from their competition through the level of satisfaction received by the customer*". The relation between manufacturer and customer does not end any more at the time of sale. Levitt (1983) compares the relationship between buyers and sellers with a marriage: "*The sale merely consummates the courtship, at which point the marriage begins. How good the marriage is depends on how the seller manages the relationship. The quality of the marriage determines whether there will be continued or expanded business, or troubles and divorce*". This comparison illustrates the fact that After Sales Service has become a competitive weapon for many companies and presents a possibility to make a difference.

An important aspect of these After Sales activities is the servicing or maintenance of technical installations purchased by the customer. These customers are often companies that use these installations for their own production processes. We are therefore mainly interested in business-to-business situations. Defective parts that cause the failure of the technical system in operation, need to be replaced quickly by serviceable parts in order to guarantee the continuity of the production process of the customer. The manufacturer usually agrees in service contracts to guarantee a certain service performance in case of failure of the technical system at the customer's site. This service performance is often measured in terms of responsiveness to customer's calls: how fast is a service

engineer at the customer's site to solve the customer's problem? Two important characteristics complicate the process of solving the customer's problem. First, the customers, and therefore also the installed base of technical installations, are usually scattered over a large geographical area. Many international companies have, for example, a central European organization for the supply of service parts to customers throughout Europe. This geographical dispersion of customers necessitates the use of a complex multi-echelon distribution network. Second, the assortment of parts that is stocked for servicing activities can be enormous. More than one hundred thousand different part numbers in stock is no exception. In most cases it is economically infeasible to stock these different parts at every location in the distribution network. The question therefore arises how to determine the parts assortment in the different stocking locations in the network. These two characteristics complicate the process of getting the right service part at the right time at the right customer. Gross *et al.* (1981) addressed this design problem of multi-echelon systems for consumer goods. In this thesis we focus on the design and control of multi-echelon systems for service parts. We explicitly mention *design* and *control* because we believe that these issues are strongly related. Too often these issues are considered as independent phases: first the design question is answered and, given this design, the operational control of the primary processes is optimized. However, design decisions that do not take into account operational control issues can have severe (financial) consequences in the operational phase. We therefore advocate an approach in which both issues, design and control, are simultaneously considered.

Several developments have increased the pressure on the service parts business in recent years:

- Economical situation
- Higher product reliability
- "Free trade" competition

First, economic developments in the late eighties and in the beginning of the nineties forced companies to reduce costs and to improve service. The inventories of service parts had to be reduced and at the same time the service to the customer had to be increased. Effective and efficient control of inventories of service parts becomes more and more critical. However, this control of service part inventories is complicated due to lack of reliable (demand) data. Second, higher product reliability also increases the pressure on the service parts business. Due to a higher reliability of the initial product, the demand for service parts decreases. Third, the competition of the "free trade" threatens the service parts business of manufacturers. The service parts with a high turnover are very attractive for other suppliers in the market. For these parts the competition is very strong. The manufacturer, however, also has to supply the service parts which are rarely needed and have a high risk of obsolescence. Therefore, the most profitable parts (with a high turnover) are submitted to strong competition whereas the less profitable parts (with no or very low turnover) are left to the

original equipment manufacturer.

The situation described above motivated this study. Effective and efficient control of service part inventories is essential to many companies nowadays. In the Netherlands this has led to the foundation of the Parts Business Forum: a forum in which managers from various companies, scientific researchers, and consultants meet each other on a regular basis to discuss issues of present interest in the field of service part logistics. These discussions have stimulated this research and have provided an excellent practical frame of reference (see Appendix A for more details).

In this thesis we present a general framework, a so-called Service Part Supply System (SPSS), for the allocation and control of service parts in complex distribution networks. We present several flexibility options that can be used to increase the cost and service performance of the system. For each flexibility option we discuss the associated economic trade off and the interaction with other flexibility options in the network.

1.2 Service Part Supply System: an example

Before we introduce a theoretical framework for a general Service Part Supply System, we present an arbitrary example of such a system in practice. A manufacturer of copiers produces these systems in many varieties at a production plant in the Netherlands. The copiers are primarily leased to customers all over Europe and the manufacturer is responsible for the operational availability of the copiers at the customer's site. This responsibility implies that whenever a copier fails at a customer location, a service engineer must visit the customer within an agreed time span to restore the failed copier. In order to realize this service performance for all customers, the company makes use of an extensive European network of inventory locations and transportation links. Figure 1.1 shows a simplified representation of the network.

A central warehouse, located in the Netherlands, stocks an assortment of service parts of approximately 30.000 different part numbers. These parts are supplied from three different sources: external vendors, the production plant, and repair centers. The central warehouse acts as a supplier to the national warehouses in the different European countries. These national warehouses in their turn act as a supplier to the regional warehouses situated in each country. Finally, from these regional warehouses the service engineers pick up service parts needed for solving the customer's problem. Next to this option, the service engineers have a small assortment of service parts in their cars, the so-called car stocks. The distribution network described here consists of four echelons: central level, national level, regional level, and car stock level. Next to the normal replenishment channels described above, the company can make use of some emergency replenishment channels

when needed. These emergency replenishments usually consist of direct deliveries of certain parts by courier services, from one stocking location in the network to another stocking location.

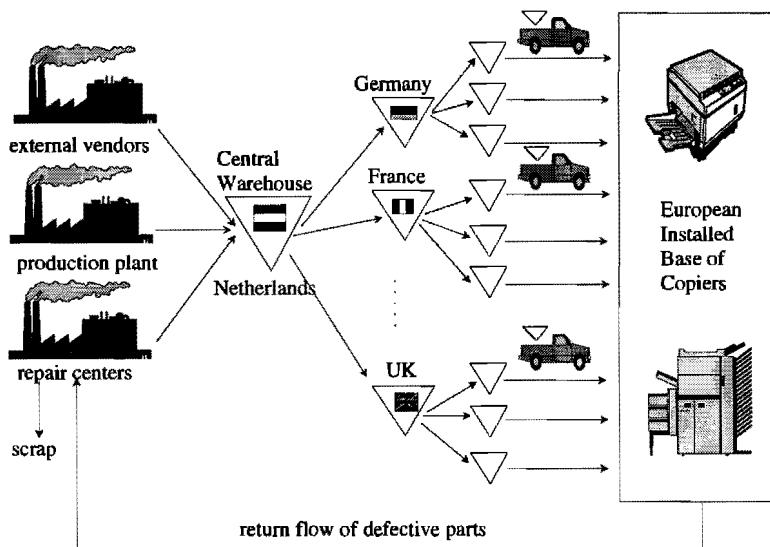


Figure 1.1: Practical example of a Service Part Supply System

The service engineer replaces failed parts by serviceable parts. The failed parts are collected by the service engineer and returned to the regional warehouses. Some of these parts are technically and economically repairable. These parts, the so-called repairables, are repaired at specialized repair centers. After repair they are sent to the central warehouse and can be used again as service parts. Therefore, these repairables form a closed loop flow of parts through the network. However, since parts are not repairable forever, they will eventually leave the system. Failed parts that are not repairable either technically or economically, the so-called consumables, are scrapped. Materials and components of these scrapped parts are often recycled and reused for other purposes.

1.3 Characteristics of service part logistics

Operational control of service part flows in distribution networks differs significantly from operational control of finished products in distribution networks. In previous research the author of this thesis addressed the latter topic (Verrijdt and De Kok, 1995, 1996, Lagodimos, De Kok and Verrijdt, 1995). In this thesis we concentrate on service part logistics. In this section we will discuss

Introduction

the following five characteristics that distinguish service part control from general production and inventory control concepts:

- Service part demand (Section 1.3.1)
- Service part life cycle (Section 1.3.2)
- Service part cost factors (Section 1.3.3)
- Service part performance measures (Section 1.3.4)
- Service part control issues (Section 1.3.5)

1.3.1 Service part demand

Demand for service parts originates when technical systems installed at customer sites fail or are in need of (planned) maintenance activities. In order to repair the failed system or carry out the specified maintenance activity, specific service parts are needed. In case of scheduled maintenance activities such as overhaul it is possible to predict the demand for service parts to a certain level. However, when dealing with unexpected failures of technical systems, the kind of parts needed and the number of parts needed is unknown beforehand. It is this second type of demand that we consider primarily in this dissertation. The timing and the quantity of demand is therefore highly unpredictable, since it is caused by random failures of technical equipment. The failures of technical systems therefore determines the demand process of service parts. A distinction has to be made between mechanical equipment and electronic equipment. Failures of mechanical systems can be predicted more accurately than failures of electronic systems, since it is based on physical observations such as wear out, number of running hours, etc. Condition monitoring plays an important role in this field (see e.g. Mann, 1983). However, more and more technical systems comprise electronic parts that have a highly unpredictable failure behavior. Failures of such systems occur randomly over time and various types of electronic parts can be the cause of the failures. The demand for service parts is therefore also highly unpredictable in terms of timing (when is a part needed?) and quantity (what part is needed?).

Fast moving / Slow moving

Most technical systems consist of thousands of different parts. Failure of a system is often caused by one specific part. A Pareto analysis of the annual demand per part number generally shows that a small fraction of the assortment (i.e., less than 5 %) is responsible for a large fraction of the annual turnover (i.e., more than 80 %) in service parts. The remaining portion of the assortment hardly contributes to the turnover but has to be kept in stock just in case they are needed. Some of the part numbers will never be demanded during the product life cycle. A service part that has an annual demand of five or more, for example, is already considered a *fast moving* item in the parts

business. A *slow moving* item on the other hand is a service part that has a demand of once every five years or even less. And these slow moving parts make up the greater part of the assortment. It is this typical demand characteristic of the assortment that distinguishes service parts from finished products. Figure 1.2 shows a typical example of such a Pareto analysis of service parts inventories in terms of annual demand and turnover (source: Parts Business Forum).

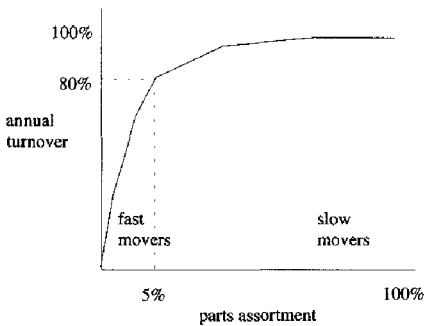


Figure 1.2: Characteristic Pareto analysis of service part inventories

Usage / Consumption

A typical phenomenon of service part demand is the distinction between *usage* of service parts and *consumption* of service parts. Service parts are needed by service engineers to solve customer problems. When a machine fails, a customer often contacts a help desk and communicates the problem as good as possible. The help desk then contacts a service engineer and sends him or her to the specific customer. From the problem description given by the customer, a first diagnosis is made and the expected parts needed to solve the problem are identified. These parts are sent to the customer's site. Sometimes the engineer picks up these parts in a local warehouse nearby. Not all parts that are sent to the customer are needed to solve the customer's problem. The parts that are not needed to solve the problem (or parts that were needed only for diagnostic purposes) are returned to the warehouse. These parts are only used for the service activities of the engineer but were not really consumed. This distinction between usage and consumption of service parts is important when modelling the demand process. Most models do not explicitly incorporate the return flow of service parts that were not consumed by the service engineer. When using *usage data* in these models, the service performance is *underestimated* since a fraction of the parts returns unused to stock. When using *consumption data* in these models, the service performance is *overestimated* since the actual demand for service parts by service engineers is higher. Models that do not

explicitly incorporate the return flow of unused parts should be applied with care to practical situations in which there exists a significant difference between usage and consumption data.

1.3.2 Service part life cycle

Service parts are used to support a technical system. Most of these technical systems have a traditional product life cycle that consists of the following four phases (see e.g. Stahl and Grigsby, 1992):

- 1) *Introduction phase*: The technical system is introduced in the market. The system still has to gain a solid position in the market. This phase is characterized as a push phase: the product has to be pushed into the market.
- 2) *Growing phase*: The technical system has gained a solid position in the market and demand is increasing. This phase is characterized as a pull phase: the market pulls the product from the manufacturer.
- 3) *Saturation phase*: The demand for the technical system becomes stable and competitors introduce competitive products. Competition is intensified and the emphasis in this phase is on cost efficiency.
- 4) *Decline phase*: The attractiveness of the product decreases and demand decreases as well. The production of the technical system is reduced and the decision is finally taken to terminate the production.

Figure 1.3 shows the development of the turnover of the product in the consecutive phases of the life cycle.

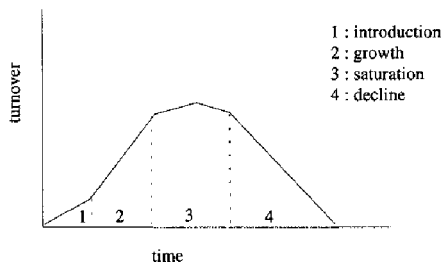


Figure 1.3: Product life cycle

The service parts that are needed to support the technical system under consideration also have a life cycle: the service part life cycle. We define the following three phases in the life cycle of service parts (Fortuin, 1980):

- 1) *Initialization phase*: A new product is introduced in the market and service parts are needed to support this product. It is very important to ensure a high service performance for these first products that are introduced in the market. The main decision to be made in this phase is what service parts to stock and how much. This phase corresponds to the introduction phase of the product life cycle.
- 2) *Normal phase*: Once a product is introduced in the market, historical data about realized demand becomes available and can be used to forecast future demand and to calculate required stock levels. This phase corresponds to the remaining production phase of the technical system (growth phase, saturation phase, and decline phase).
- 3) *End phase*: The end phase starts when the production of the technical system is discontinued. The manufacturer of the technical system has the obligation to provide service parts for a certain number of years to support the installed base of these systems in the field. This service period can be very long.

Figure 1.4 shows the development of the turnover of service parts in the consecutive phases of the life cycle.

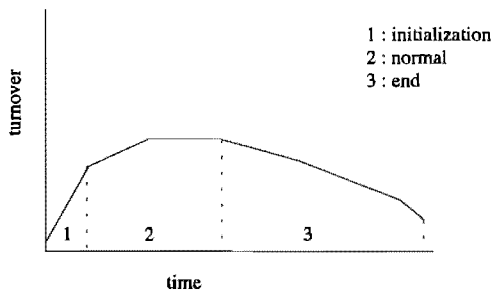


Figure 1.4: Service part life cycle

The most difficult phases from an (inventory) control point of view are the initialization phase and the end phase. Some models have been developed that address the issue of inventory control in these phases (see e.g. Fortuin 1980, 1981, 1984, Teunter and Klein Haneveld 1995, Teunter and Fortuin, 1996). In the initialization phase stock levels have to be determined for all service parts in all inventory locations in the distribution network. However, since the product has been newly introduced in the market, no historical demand data is available to forecast the expected future

Introduction

demand for service parts. The only source of information that can be used to forecast the demand for service parts is test data obtained from the R&D department. However, in many cases this information is either not available or it is available but not in such a form that it can be used to forecast future demand. As a result, the inventory locations are usually overstocked in the initialization phase to be on the safe side. The difficulty in the end phase lies in the fact that discontinuation of the production of the technical system often also implies the discontinuation of the production of the associated service parts. Maybe it is possible to produce service parts after discontinuation of production, but than only at high cost. The 'final order' for service parts when production is discontinued, is therefore very important. However, it is very difficult to forecast the demand for service parts over such a long service period. In aircraft industry, for example, the manufacturer has an obligation to provide service parts thirty years after discontinuation of production of the airplane. In this thesis we consider service parts that are either in the normal phase (in which supply is always possible) or in the end phase (in which there is enough inventory to satisfy any future demand).

The average demand for service parts often changes during the different phases of the service part life cycle. An important cause of this is the aging of technical installations. Other circumstances can also cause a non-stationary demand process over time. Think for example of the demand for service parts for military aircraft in peace time or in war time. There are different techniques that can be used to model demand processes with a mean that varies over time, such as Bayesian statistics and James-Stein statistics. We refer to Sherbrooke (1992b) for a detailed analysis of these techniques.

1.3.3 Service part cost factors

There are several cost factors associated with the supply of service parts in a distribution network. A distinction can be made between costs related to structural measures and costs related to operational measures (Corbey, 1995). Wouters (1993) refers to these different costs as full costs and relevant costs. Structural measures are characterized by a long implementation time (i.e. time needed to implement the measure in the organization), a long commitment time (i.e. minimum time period in which the measure is applied in the organization), and a high economic risk (i.e. high investment is required to implement the measure). Operational decisions on the other hand are characterized by a relatively short implementation time, a relatively short commitment time, and a relatively low economic risk. With respect to the supply of service parts in a distribution network, structural costs are related to long term (e.g. more than five years) investment decisions, such as number, size and location of warehouses and investment in Information Technology. Operational costs are related to the daily operational processes that take place, such as purchasing parts, replenishing inventory, and applying emergency procedures. In this thesis we restrict ourselves to these operational costs: costs that are related to the time period in which the policy decisions are implemented. When analyzing

trade-offs for implementation of flexibility in a Service Part Supply System, we should only consider those costs that are affected by the decision whether or not to implement flexibility. As Wouters (1993) points out, in practice too many operational (short term) decisions are based on structural costs (or full costs) that are not influenced by this decision.

When considering the use of flexibility in a Service Part Supply System, one has to make an economic trade-off between the relevant (i.e. operational) costs that are affected by this decision. We identify the following four main operational cost factors:

- *Inventory holding cost*
- *Normal replenishment cost*
- *Emergency replenishment cost*
- *Shortage penalty cost*

Cost for holding inventory at warehouses throughout the distribution network is usually the most critical cost factor. The main goal of using flexibility is therefore in most situations to reduce inventory and therefore also reduce inventory holding cost. Inventory holding cost is expressed as a percentage of the *purchase price* of a service part. Note that we explicitly mention the purchase price of a service part and not the selling price. This is because reduction of inventory leads to cost savings by reducing the acquisition of new parts (at purchase price!) at vendors. This holding cost percentage is a composition of several percentages: interest, space, and obsolescence. The interest percentage represents the opportunity cost of the capital that is tied up in the inventory, the space percentage represents the cost of renting storage capacity, and the obsolescence percentage represents the risk that parts become obsolete in the course of time. Space cost may only be considered in the economic trade-off if the storage capacity (that becomes available due to inventory reduction) can be used for other purposes as well (e.g. public warehousing). If this is not the case, space cost should be eliminated from the trade-off, since this cost factor is not influenced by the decision to implement flexibility. Obsolescence cost is a very important holding cost factor, especially for service parts with a low demand. This cost factor represents the probability that a service part will become obsolete in the future (either due to aging or due to technological innovation) and has to be replaced by a new part. The obsolescence percentage of a service part (expressed as a percentage of the purchase price) is usually determined by the level of technology that is used in producing the part. This percentage is likely to increase in the course of time.

Normal replenishment cost is incurred for the normal ordering process within the distribution network (i.e. between warehouses) according to the specified inventory replenishment policies. The most important cost factor for normal replenishments is the transportation cost. Many companies for instance have contracts with third party service providers who take care of the routine shipments

Introduction

from the central European warehouse to the national warehouses in Europe. Distribution of service parts within a country is often done by local service providers. Other cost factors that are related to normal replenishments, such as administrative cost for order handling and material handling cost in the warehouses, are less important. Moreover, these costs are often fixed for a longer period and should therefore not be considered in the economic trade-off for implementing operational measures.

Emergency replenishment cost is incurred for using emergency supply flexibility in the distribution network. Most companies have different kinds of emergency options available, with different associated costs. Emergency requests for service parts usually originate when a technical system of a customer has failed, and the service engineer is in need of a specific service part. In order to minimize the down time of the technical system, that specific service part has to be supplied from some stocking location in the network as soon as possible. This means that a special direct delivery for one part has to be made from anywhere in the network to that specific service engineer. The cost of using such a direct delivery is usually the cost of using a courier service such as DHL and UPS. Emergency replenishment cost can also be incurred when the part that is needed for solving the customer's problem is not in stock anywhere in the system and has to be manufactured. Next to the emergency transportation cost (to get the part from the production plant to the customer's site) extra emergency production cost is incurred for producing this specific part. Such an emergency production order can interfere with the normal production activities and can result in high extra cost.

The last relevant cost factor is the penalty cost for not having service parts immediately available to carry out repair or maintenance activities. The (financial) consequences of not having service parts available to repair the technical system at a customer site strongly depends on the circumstances. The situation where an airplane is grounded somewhere in the world or an oil drill installation in the North Sea is not in operation because the needed service parts are not available, is quite different from the situation in an office building where a copier is out of order. The consequences of a technical system being down ranges from inconvenience and annoyance for the customer to production losses in the magnitude of millions of dollars. Sometimes it is contractually agreed that a fixed penalty cost is paid to the customer for every unit of time that the system is down. These costs are real costs. However, in many situations in practice it is very difficult to put a price tag on the absence of service parts since it represents customer dissatisfaction and loss of goodwill. In that case the incurred cost is virtual and subjective. The impact of a stock-out of one service part type can also differ per situation in terms of penalty cost. Dekker *et al.* (1996) developed a model in which some of the stock is reserved for critical demand (i.e. with high penalty cost). Once the inventory on hand hits a critical level, only critical demand is satisfied from stock on hand. Non-critical demand (i.e. with low penalty cost) is backordered until the inventory on hand exceeds the critical level again.

1.3.4 Service part performance measures

A very important issue in service part management is the measurement of the performance of supplying service parts to customers. Customer satisfaction is very important to ensure continuation of business. Lack of performance can also have a direct impact on revenues due to explicit contractual penalty costs. Companies use a wide variety of qualitative and quantitative performance measurement tools to measure the service performance within the distribution network (e.g. from central warehouse to national warehouses) and towards the external customer. Cohen and Lee (1990) mention five service measures commonly used in practice:

- 1) *Part Unit Fill Rate*: The fraction of demand (either usage or consumption) of a service part delivered from inventory on hand from some stocking location over some period of time. This measure is part specific but can also be calculated for the aggregate demand at some stocking location.
- 2) *Part Dollar Fill Rate*: This measure is a modification of the Part Unit Fill Rate in that it measures the fill rate in terms of money value instead of quantities.
- 3) *Order Fill Rate*: The fraction of internal replenishment orders (consisting of a number of different service parts) that can be completely filled from inventory on hand. This measure is important when the fixed cost per order is high or partial order filling is not allowed.
- 4) *Repair Order Completion Rate*: Fraction of repair jobs carried out at the customer that is not delayed by part shortages. Since repair jobs can only be carried out when all parts needed are available, this measure is typically lower than the individual part fill rates.
- 5) *Customer Delay Time*: Time period between the identification of a service need and the meeting of that need. This measure is closely linked to the previous measure: if some parts are not available for the repair job, alternative sources have to be used to minimize the delay.

The first three measures are encountered most frequently in practice because they measure the internal service performance (per orderline or per order) within the distribution network. The latter two measures reflect the service performance as perceived by the customers. Although these measures are very important, they are less frequently used in practice because the required information technology is not always available. The customer's perception of service is often also measured through qualitative customer surveys (e.g. by phone). In service contracts with customers the service performance is often described in terms of response time (or customer delay time) restrictions, e.g.: "90% of all calls for service will be attended to within 8 hours, and 99% of all calls will be attended to within 24 hours". Such a performance is measured periodically, e.g. per month or per year.

Introduction

In addition to the distinction between internal oriented service measures (between warehouses within the distribution network) and external oriented service measures (service perceived by the customer), it is possible to make another distinction:

- *quantity-weighed*
- *time-weighed*

Measures 1 to 4 are typical examples of quantity-weighed measures in that they measure the fraction of 'transactions' that is executed without delay. However, they do not measure the delay that is incurred for those 'transactions' that are executed with some delay. A fill rate of 98% seems very well, but it does not say anything about the remaining two percent of demand that is not satisfied from stock on hand. The fill rate measure does not reflect whether it takes one hour or one week to meet this demand. However, time-weighed measures, such as measure 5, reflect the speed with which service activities are carried out. This can be very important due to contractually agreed penalty cost in case the response time exceeds a certain threshold value. The Customer Delay Time or response time is strongly influenced by the flexibility that exist in the distribution network to use alternative sources. Both quantity-based measures and time-weighed measures are important and they should be used together.

An example of a service measure that combines both these aspects (quantity and time) and that is often used in the literature is the *average number of backorders* at a stocking location: total number of days that backorders (of service parts that are needed to restore a failure) are outstanding divided by the length of the time period under consideration. Note that this service measure weighs one backorder with a duration of e.g. ten days equal to ten backorders each lasting one day. This service measure accounts for both quantity of shortages and duration of shortages. Another advantage of this measure is that it can be shown that minimizing the expected number of backorders is almost equivalent to maximizing the availability of the technical system under consideration (Sherbrooke, 1992b, p. 38).

1.3.5 Service part control issues

A distinction has to be made between network design and control of service part flows for a given network design. There is however a very strong linkage between these two issues and one can not be considered without the other. Network design issues refer to strategic decisions such as the number of warehouses in the network, the capacity of these warehouses, the depth of the network (how many echelons?), linkages between warehouses, and allocation of customers to warehouses. For a given network design the following control issues have to be dealt with:

- 1) What are the criteria with respect to cost and service performance?
- 2) How to determine the assortment of service parts for every inventory location in the network (the so-called 0-1 decision)?
- 3) What inventory control policies must be used for different categories of service parts?
- 4) What emergency sourcing policies must be used for the different categories of service parts in case of a stock-out situation?
- 5) What are the control parameters for the decision problems described above?

These five control issues are a mix of strategic, tactical and operational decisions. The performance criteria with respect to cost and service have already been addressed in Sections 1.3.3 and 1.3.4. The 0-1 decision concerning the composition of the parts assortment at different locations is a strategic issue in the service parts business. This 0-1 decision is a typical trade-off between service and cost. Positioning inventory at a low level in the network (i.e. near the customer) ensures a high service performance towards the customer but also a high investment in inventory, since there is a large number of local stocking locations. Positioning inventory at a high level in the network (i.e. only at the central warehouse) ensures a low investment in inventory but also a low service performance towards the customer, since the customer is in general located at a larger distance. Distribution costs are also higher in this situation, because parts are distributed every time a demand takes place.

Given the assortment of service parts for a stocking location, the inventory control policy has to be determined. The choice of inventory control policy depends primarily on price and demand characteristics of service parts, but also on the criticality of the service part for the functioning of the technical system. Since the majority of service parts are slow moving and expensive items, the one-for-one replenishment policy (or $(S-1,S)$ -policy) is by far the one most frequently used in practice. However, for fast moving and/or less expensive service parts ("the nuts and bolts") other policies are implemented that allow for order quantities larger than one. This thesis concentrates on the first kind of service parts that are controlled by one-for-one replenishment policies.

Next it has to be determined from where the inventories in a stocking location are to be replenished and in which mode. In other words, which inventory location supplies which inventory location? An important distinction that has to be made here is the distinction between normal replenishments and emergency replenishments. For normal replenishment of inventory, procedures are determined how, when, and where to order. However, in case of an emergency (a customer is waiting for a specific service part to solve a problem) more than one option may exist to procure that specific service part. Examples of such options are direct deliveries (or repairs) from a higher echelon and lateral transshipments between locations at the same echelon.

Finally, once an appropriate policy for normal and emergency replenishments has been chosen, the control parameters have to be established. This typically concerns the calculation of inventory reorder levels and order quantities. Many models in Operations Research have contributed to the determination of these control parameters in distribution environments.

The models presented in this thesis can be used to optimize these operational control issues of the primary processes in Service Part Supply Systems. However, these models can also be used to evaluate different network designs and evaluate their impact on operational costs. This proves again the strong relation between design and control of Service Part Supply Systems.

1.4 Objective of this thesis

The design and control of distribution networks for service parts is the subject of this dissertation. The emphasis that is put nowadays in industrial practice on the improvement of the service performance towards the customers and the reduction of the systemwide inventory motivated us to conduct this research. Our main objective was to investigate the benefits that can be expected from using different kinds of flexibility in distribution networks for service parts. How does the use of flexibility in such systems contribute to the goal of increased service performance and reduced cost?

The added value of this thesis to the state of the art research in this field can be summarized as follows:

- Presentation of a flexibility framework for service part logistics.
- Development of two analytical models that can be used to investigate the trade-off between cost and service performance for some specific flexibility policies.
- Development of a simulation model that can be used to evaluate and compare different flexibility policies in a multi-echelon distribution network.
- Increase the insight in the complex relations between network structure and flexibility policies.

We present a framework for the control of service parts in a distribution environment, the so-called Service Part Supply System (SPSS), in which we identify a number of flexibility options that can be implemented in order to increase service and decrease cost. This framework is the first to present a global overview of the different kinds of flexibilities that can be used for the control of service parts. The framework is applicable for both repairables and consumables and a distinction is made between *repair flexibility* (referring to the repair processes of failed service parts) and *supply flexibility* (referring to the supply processes of service parts in a distribution network).

Two analytical models are developed that can be used to quantify the trade-off between cost and service performance. The first model, the Emergency Repair Model, is developed to investigate the use of emergency repair flexibility. How much reduction in inventory can be obtained when you can make use of a fast (and expensive) emergency repair procedure? The second model, the Emergency Supply Model, is developed to investigate the use of lateral transshipment and direct delivery flexibility in a two-echelon distribution system. This kind of flexibility can be very beneficial in case of high penalty cost. These two analytical models can be used as tools to investigate economic trade-offs for some specific flexibility policies.

To evaluate more complex flexibility policies in a distribution network we developed a simulation tool. Policies that consist of different flexibility modes (e.g. lateral transshipment flexibility, direct delivery flexibility, pipeline flexibility) can be evaluated using this tool. In this way we are able to quantify the impact on cost and service performance of individual flexibility modes.

The two analytical models and the simulation model increase the insight in the complex relations that exist between network structure and flexibility policies. Numerical evaluation of these models help us to find practical guidelines for when to use what kind of flexibility in distribution systems for service parts. We can formulate the main research question as follows:

In what industrial environment, under what conditions, does what kind of flexibility contribute most to the performance of an SPSS?

The industrial environment plays an important role when considering the logistic control of the flow of service parts in an SPSS. A manufacturer of airplanes who is responsible for the supply of expensive service parts (e.g. rotor blades with a value of thousands of dollars) to a few customers (e.g. twenty airlines worldwide), has a different logistical control structure than a manufacturer of personal computers who is responsible for the supply of relatively cheap service parts (e.g. a key of a keyboard) to many customers (e.g. hundreds of retailers). The characteristics of the technical system that is produced and the characteristics of the manufacturer influences the design and control of the Service Part Supply System.

The conditions under which the service part flow has to be controlled can differ as well. The variety of strategic goals that can be formulated for service part logistics is quite large and these goals can be conflicting with each other. For example, the strategic goal "minimize the total cost of operating the supply chain" can be conflicting with the strategic goal "maximize the fill rate for every stocking location". The condition under which the Service Part Supply System must operate therefore influences the expected benefits from using supply flexibility.

Introduction

The framework, the analytical models, and the simulation tool help us to answer this fundamental question: when should we use what kind of flexibility? In this thesis we consider the use of flexibility options as an integral part of the control and design of service part distribution networks. Where most research is focused on improving the performance of routine activities (e.g. inventory replenishment) in distribution networks for service parts, we consider the performance of non-routine activities (e.g. emergency replenishment in case of stock-out situations) as a fundamental part in the design and control of service part distribution networks.

Research methodology

We started our research with a thorough review of the existing literature. Based on this literature review we identified several flexibility options that can be applied for service part logistics. This resulted in the development of the Service Part Supply System framework. Five case studies that were carried out during the last four years are positioned within this framework. Next we developed two new analytical models to model flexibility options explicitly. Again, these new models are developed in line with the existing literature. Validation of both models was established by means of simulation. To extend the research to more complex flexibility policies we developed a simulation tool. The experimental design that was used for the simulation study is based on fractional factorial R-3 design. Finally, we reviewed our research efforts with respect to some practical issues that were discussed in a forum of companies active in the service parts business.

1.5 Outline of this thesis

This thesis is organized as follows. In Chapter 2 we present an extensive overview of the existing literature in the field of service part inventory control. In Chapter 3 we introduce the Service Part Supply System, a framework for the control of service part flows in a complex distribution and repair network. In this framework we identify and discuss the use of different flexibility options that can be applied to improve the performance of the system as a whole. In Chapter 4 we address one specific flexibility option: the trade-off between emergency repair and inventory investment in service parts. For this purpose we developed the Emergency Repair Model. In Chapter 5 we discuss a second specific flexibility option: the use of emergency lateral transshipments between inventory locations at the lowest echelon and direct deliveries from inventory locations at a higher echelon in distribution networks. For this purpose we developed the Emergency Supply Model. In Chapter 6 we present a simulation model that we use to evaluate the impact of more complex flexibility policies in a Service Part Supply System. This enables us to investigate the impact of separate flexibility options on cost and service performance. Finally, in Chapter 7 we summarize our main conclusions and make some recommendations for further research.

Chapter 2

Literature review

2.1 Introduction

In this chapter we present an extensive literature review of the research on the control of inventory positions and repair capacity in Service Part Supply Systems. A distinction can be made between research focused on determining optimal stock allocation in the supply chain, and research focused on simultaneously determining the optimal repair structure (in case of repairable items) and the optimal stock allocation in the supply chain. The first category of literature assumes that in case of repairable items the repair structure, i.e. the location of repair centers in the supply system and the capacity of these repair centers, is given and tries to determine the optimal allocation of service part inventories in the supply chain. The second category of literature on the other hand considers the repair structure as a decision variable as well (e.g. number and location of repair centers or expensive test equipment) and tries to determine simultaneously the optimal inventory allocation of service parts and the optimal allocation of repair capacity in the supply system. The research presented in this thesis is closely related to the literature in the first category and therefore we primarily focus our attention to this category (Section 2.2, 2.3 and 2.4). A short review of the second category of literature is given at the end of this chapter (Section 2.5).

In Section 2.2 we start with describing the METRIC model (Sherbrooke, 1968) which is widely considered to be the first model to capture the different interactions in multi-echelon inventory systems for service parts. Several extensions to the original METRIC model have been developed since which are described in Section 2.3. A separate class of inventory oriented supply chain models for service parts was developed by Cohen *et al.* These models are described in Section 2.4. A short review of the literature that considers the repair structure as a decision variable is presented in Section 2.5. Finally, in Section 2.6 we discuss the literature and position the research presented in this thesis in relation to the existing literature.

2.2 The METRIC model

When looking at multi-echelon supply systems for repairable service parts, 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 optimal inventory levels for service parts. The model was successfully implemented at the US Air Force for

calculating optimal stock levels for expensive and slow moving repairable service parts. Later on a METRIC-based model was implemented at the US Navy as well (Clark, 1981). METRIC is an approximate multi-item multi-echelon inventory model that describes the supply and repair processes of repairable service parts in a two echelon system consisting of a central depot and a given number of local warehouses or bases. The operational processes of supply and repair of service parts are described as follows. External demand for service parts originates at local level due to failures of technical systems in the installed base. In the case of the US Air Force service parts were needed for maintenance and repair activities for airplanes stationed at different air bases in the USA. The defective parts are immediately replaced by serviceable parts from stock on hand if available, otherwise the demand is backordered. It is assumed that the probabilities that a defective part is repaired at the local base (in case of relatively simple repair activities) or at the central depot (in case of relatively complex repair activities) are known for all parts. If the part is sent to the central depot for repair, a replenishment order is simultaneously generated to replenish the local inventory at the specific local base. If the central depot has serviceable stock on hand a service part is immediately shipped to the local base, otherwise the demand is backordered and satisfied when parts become available from the repair shop at the central depot. Backordered demand of the different bases at the central depot is satisfied using the First Come First Serve policy at the central depot. Note that the situation described above implies that all inventory locations in the system (local bases as well as central depot) apply a one-for-one inventory replenishment policy (or $(S-1, S)$ policy with stock level S). The goal of the METRIC model is to determine the optimal stock levels of all service parts at all inventory locations in the system such that the sum of the expected backorders of all service parts at *all local bases* is minimized. The constraint is that a given budget is available for investment in service part inventory.

A number of important assumptions are made in the formulation of the METRIC model. The demand for service parts at the local bases is assumed to be (compound) Poisson distributed. Furthermore it is assumed that all parts can be repaired an infinite number of times. As a result, external procurement of service parts is not allowed. However, the most important assumption is related to the modelling of the repair processes. In METRIC it is assumed that the repair times for all defective parts (both at the local bases and at the central depot) are independent and identically distributed with a given mean. Defective parts are immediately taken into repair and do not have to wait for repair capacity to become available. This *infinite repair capacity* assumption allows the application of Palm's Theorem (Palm, 1938) which states the following:

If defective parts are generated by a stationary Poisson process and repair times are independent and identically distributed random variables, then the number of parts undergoing repair in steady-state is Poisson distributed with a mean equal to the product of the failure rate and the mean repair time.

This important assumption makes it possible to formulate closed form expressions for all relevant service and cost performance measures. 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 demand 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. In this extension it is assumed that all defective parts in a demand 'cluster' have the same repair time.

2.3 Extensions to the METRIC model

The METRIC model from 1968 served as inspiration to many scientific researchers in the decades that followed. The first important extension to the original METRIC model was developed by the initial author himself. It concerns the recognition of a hierarchical product structure, also known as multi-indenture structure. In the case of the US Air Force the METRIC model calculated optimal stock levels for all service parts at all stocking locations in the system. No distinction was made between end-items (i.e. plane engines), modules (i.e. subassemblies that make up a plane engine), and components (parts that make up a module). The original METRIC model minimized the total expected backorders of *all* these items at the local bases. In practice, however, only shortages of end items (i.e. plane engines) immediately affect the down time of the technical systems that are supported (i.e. planes). A shortage of modules or components has no direct impact on the down time of the technical system itself. It only affects the down time indirectly, because the repair of the end-item is delayed. Sherbrooke (1971) was the first to recognize this multi-indenture relationship in the product structure. He develops an expression for the operational availability of a two-level technical system (i.e. military aircraft) that is supported by stocking end-items (e.g. engines) and modules (e.g. engine subassemblies). The operational availability criterion is identical to the expected backorder criterion in the METRIC model. Failure of an end-item is assumed to be caused by the failure of one module or an external cause. Again the important assumption is made that checkout times of end-items and repair times of modules are independent and identically distributed such that no queuing occurs (i.e. infinite capacity is assumed). The model is evaluative in nature and is developed for a single base situation. Muckstadt (1973) implemented the multi-indenture product structure in the original METRIC model. This model, called MOD-METRIC, determines for a given budget the optimal allocation of end-items and their comprising modules in a two-echelon system, such that the expected backorders of *end-items* at all local bases is minimized. It is assumed that failure of an end-item is caused by exactly one module. The model was implemented by the US Air Force for the F-15 weapon system. An extension of the MOD-METRIC model to a three-echelon situation is described in a later paper by Muckstadt (1979).

An improvement of the METRIC model from a computational point of view is presented by Muckstadt (1978). The METRIC optimization procedure for finding the optimal stock levels can be very time-consuming for many practical situations with a sizable system stock level. Muckstadt develops an alternative approach for finding the optimal stock levels that shows a reduction in computation time of nearly 50 percent for two hypothetical cases. He also presents an approximation method for estimating the optimal depot stock level that reduces the computation time considerably.

The METRIC model is completely conservative. Failed parts can be repaired an infinite number of times, either at the local bases or at the central depot. Simon (1971) extends the METRIC model to allow for positive condemnation rates: a failed part that arrives at a local base is repaired at that base, condemned at that base, or sent to the central depot for repair. Note that a condemnation rate equal to zero reduces the model to the original METRIC model. A condemnation rate equal to one reduces the model to a situation where all parts are consumable. The inventory manager at the central depot applies an (s, S) inventory policy: when the inventory position at the central depot drops below s (due to condemnation), an external procurement order is issued to increase the inventory level up to S . Because all bases apply a one-for-one ordering policy, the depot effectively applies an (s, Q) ordering policy, where $Q = S - s$. Simon derives exact expressions for expected backorders, stock on hand, and parts in repair at each stocking location. Shanker (1981) presents a similar analysis for the situation where failed parts arrive in batches at the local bases. A batch inspection policy is applied that determines if an arriving batch of failed items as a whole is either base repairable, depot repairable or condemnable. Repair times are assumed to be deterministic.

In the METRIC model the allocation policy at the central depot of repaired parts to local bases with outstanding orders is the First Come First Serve policy. Miller (1974) presents a heuristic rule, the Transportation Time Look Ahead policy, that sends an item completing depot repair to the local base whose marginal decrease in expected backorders will be the greatest at x days into the future, where x represents the constant transportation time from the central depot to that specific base. Miller models the system as a Markov decision process in which all parts completing depot repair are immediately sent to the bases. This rule is shown to be optimal in a modified model.

Muckstadt and Thomas (1980) investigate the benefits of using a multi-echelon method for the control of multi-echelon inventory systems for service parts. Their model is described for consumable service parts. Instead of performing repair activities the central depot purchases new service parts at some production plant which is assumed to have infinite supply. They extend the original METRIC model with the option of direct deliveries from the central depot or the production plant in case of a stock-out situation at the local warehouse. The main conclusion is that multi-echelon systems managed by using adapted single-echelon models (which can be observed often in practice), can be dramatically inferior to models that take advantage of the system's

Literature review

structure. Hausman and Erkip (1994) use the same model and data set to demonstrate that the cost increase when using single-echelon models instead of a multi-echelon model is approximately 3% to 5% when the parameter setting is appropriately done.

The METRIC model is an approximate model by assuming that the number of outstanding orders at each base is Poisson distributed. There is however a correlation in the time delay that base replenishment orders incur when the central depot is out of stock. The Poisson distribution assumes that the variance of the pipeline stock equals the mean, which in practice is often not true. Therefore the one-moment Poisson approximation can yield significant errors when computing the 'optimal' stock levels in the system. Slay (VARI-METRIC, 1984) and Graves (1985) propose a two-moment approximation for the distribution of the number of outstanding orders at the bases. Graves models a repairable item base/depot supply system where repair is only possible at the 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) applies this two-moment approximation to the multi-indenture multi-echelon model (MOD-METRIC) and shows that significant improvements can be obtained. In fact, the performance is very close to the "true" simulation results.

The use of lateral transshipments between bases in case of a stock-out situation is not allowed in the original METRIC model. Several researchers have addressed this issue since (e.g. Lee (1987), Axsäter (1990), Sherbrooke (1992a), Dada (1992)). A detailed overview of this line of research is presented in Chapter 5 where we address this flexibility option in a service part supply system.

Cheung and Hausman (1995) address the issue of multiple failures. Repair jobs entering a repair center often consist of more than one failed part that are in need of repair. They model the situation with independent Poisson arrivals or repair jobs and a one-for-one repair policy with cannibalization. The model can be used for evaluating spares inventory allocation, part demand correlation, and parts commonality.

In conclusion we can say that the METRIC model developed in 1968 has triggered an impressive flow of papers that addresses related issues. De Haas and Verrijdt (1996), for example, apply the (MOD-)METRIC model to derive local targets with respect to service performance for individual organizational units within a multi-echelon network. A good review of the METRIC methodology in general (and some of its extensions, such as MOD-METRIC and VARI-METRIC) can be found in Sherbrooke (1992b).

2.4 Cohen models

We now focus on a series of papers by Cohen *et al.* These authors address the problem of determining optimal stocking policies for service parts in complex multi-echelon systems as well. However, unlike the research discussed in the previous sections, they assume a periodic review inventory policy for service parts. Moreover they only consider consumable service parts that are scrapped once they have failed. The goal is to minimize total cost subject to some relevant service level constraints. This line of research is very interesting since they take into account a number of practical aspects, such as emergency transshipments, demand priorities, and pooling mechanisms. The results of their research were used for developing a stocking policy model that was implemented at IBM's US after sales service logistics system with great success (Cohen *et al.*, 1990).

The first paper (Cohen *et al.*, 1986) considers a single-item single-period multi-echelon system for service parts. Demand for service parts originates at the lowest echelon of the system. 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 at a particular stockpoint to a higher echelon stocking location, it is first checked whether neighboring stocking locations, belonging to the same pooling group, have excess inventory after their demand is satisfied. If so, this excess inventory is used to meet this excess demand. 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. It is assumed that these parts are returned within the same review period in which they were demanded. The objective is to determine optimal stocking levels for each location, that minimize total expected cost per period (i.e. cost of emergency shipments, normal replenishments, and inventory holding), subject to a response time constraint (e.g. 95% of the demand is fulfilled within 4 hours). It is assumed that lead times are deterministic and that no outstanding backorders exist at the beginning of a review period. A branch and bound procedure is used to determine optimal stocking policies. An important conclusion is that the response time constraints determine the positioning of stock in the system, whereas the quantity to be stocked depends on the cost/demand class in question.

The distinction between different classes of demand (emergency requests versus normal replenishments) is further researched in Cohen *et al.* (1988). In this paper they describe a single-echelon single-item (s,S) inventory system under periodic review. The demand imposed at the stocking location consists of two priority classes: emergency demand (with high priority) and normal

Literature review

replenishment demand (with low priority). Excess demand that cannot 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 the total expected cost subject to a fill rate service constraint. Total costs consist of ordering cost, holding cost, transportation cost, and shortage cost (determined by the cost of the emergency procedure). The numerical results indicate a good performance of the heuristic which deteriorates as the fill rate constraint and lead time increase. In a later paper (Cohen *et al.*, 1992) this single-item model is extended to a multi-item situation, where a product consists of a number of different items. The service level constraint in this model is defined at product level instead of item level.

Finally, a similar single-echelon multi-item periodic review inventory system is considered in Cohen *et al.* (1989). A heuristic is presented that calculates the order-up-to-levels for all items at a stocking facility that support a particular product. The criterion is to minimize the total expected cost (i.e. ordering, holding, transportation, and shortage cost), subject to a service level constraint at product level. Two types of service constraints are considered: a chance constraint (fraction of the time that all requested 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). Excess demand is satisfied through an emergency procedure (which determines the shortage costs) and is considered lost to the stocking facility. It is assumed that the stocking facility is fully stocked at the beginning of a review period. The paper also extends the above model to a two-priority demand classification situation and to a situation where commonality between parts is allowed (i.e. a single part can be used in different end products).

2.5 Repair models

Both the METRIC based models and the Cohen models discussed in the previous sections focus on optimal inventory allocation policies for repairable or consumable service parts in distribution networks. The underlying assumption in these models is that repair capacity (in case of repairable service parts) or supply capacity (in case of consumable service parts) at the highest echelon is infinite. The repair structure, in terms of location of repair shops and capacity of the repair shops, is assumed to be given and is not considered as a decision variable. However, there is a large body of literature that considers the repair structure as a decision variable as well. In this section we give a brief overview of some of the literature that considers a trade-off between repair structure (e.g. number of repairmen in a repair shop) and inventory allocation (e.g. number of spare machines in the system). This research is mostly based on queuing theory. A more detailed overview of this kind of literature can be found in e.g. Nahmias (1981) and Cho and Parlar (1991).

Gross (1982) investigates the effect of assuming ample repair capacity (enabling the use of Palm's theorem) in a single-echelon single-item repair model with a limited number of repair channels for failed items. He compares the $M/M/\infty$ system (i.e. the METRIC model with infinite repair capacity) with an $M/M/c$ system (i.e. only c repair channels are available). Numerical results are presented that show the error that is made by 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. Ahmed *et al.* (1992) also compare infinite-source, ample-server models with finite-source, finite-server models and they show when infinite-source ample-server models can be used as a good approximation.

Scudder and Hausman (1982) use a simulation model for a repair shop with limited repair capacity to evaluate the stocking policies for multi-indentured repairable items when using different priority scheduling rules. They compare their simulation model with the MOD-METRIC model which assumes infinite repair capacity. Their results show that models assuming infinite repair capacity when it actually is not, perform quite good. In a related paper Hausman and Scudder (1982) evaluate the performance of different priority scheduling rules in such an environment in more detail. Their results indicate that dynamic scheduling rules which use inventory status information performs better than other rules. They show that the use of improved scheduling techniques can lead to an improved performance equivalent to a 20% reduction in spares inventory. Scudder (1984) investigates a similar problem where multiple failures can occur. 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. The batching problem for a repair shop with limited spares and finite capacity is addressed in a simulation study by Chua *et al.* (1993). They present a batching policy that performs very well in all the environments tested. In this policy, the size of the batch is equal to the number of failed parts of a specific type that is waiting for repair and the batch selection is based on the shortest batch processing time per part, weighed by the number of available spares.

Gross *et al.* (1983) present a model for simultaneously determining service part stock levels and the number of repair channels that minimize the total expected cost subject to a pre-specified system availability constraint. This model is an extension of a model by Mirasol (1964) that considers a single-echelon repair shop with an infinite source of demand for service parts. Gross *et al.* extend this model to a two-echelon environment (consisting of a central depot and one base) and a finite source of demand. In Gross and Miller (1984) the 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 inventory stocking is possible at both echelons. Albright and Soni (1988) analyze a two-echelon repairable-item inventory system using continuous-time Markov processes. Ebeling (1991) presents a methodology for determining the optimal allocation of repairable item inventories and repair

Literature review

channels in a system that supports an operating system that consists of different parts. Recently Abboud (1996) developed a new algorithm to compute the long-run average number of machines operating in a two-echelon repairable item inventory system.

Kaplan and Orr (1985) present a model called OATMEAL (Optimum Allocation of Test Equipment/Manpower Evaluated Against Logistics). OATMEAL determines simultaneously optimal maintenance as well as stocking 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 determines at which echelon each maintenance function will be performed, or whether the maintenance function should be eliminated. Alfredsson (1996) presents a related model in which the amount of service parts and test equipment are determined based on a life-cycle cost constraint. The model also answers the question at what level in the multi-level system the repair work should be carried out.

Daryanani and Miller (1992) evaluate the performance of a dynamic return policy of repaired items at the depot to the local bases. The repaired item is sent to the local base which has the highest number of outstanding orders. They model this situation as a single server queuing system with multiple sources. The repair times are independent and exponentially distributed and the failed items arrive from multiple sources according to independent Poisson processes. They show that dynamic return policies affect the system performance. Büyükkurt and Parlar (1993) conduct a simulation study to evaluate three different return policies in a two-echelon system with state-dependent failure and repair rates: return a repaired item to (1) the base which has the longest outstanding backorder (i.e. First Come First Serve), (2) the base where it originated from, and (3) the base which has the highest number of items at the central repair shop. This third dynamic policy proved to be superior to the two other static policies under five different optimality criteria.

Schneeweiss and Schröder (1994) present a hierarchical model for the supply and repair of service parts that was implemented at Deutsche Lufthansa AG. At the highest level the optimal number of service parts is calculated that guarantees a certain service level at minimum cost. At the lower level scheduling rules for the repair of defective parts are implemented that guarantee that the service level is actually maintained.

Finally, we address some research where the use of repair capacity in a repairable item inventory system is flexible. Most models that consider repair capacity as a decision variable concentrate on determining the optimal size of the repair capacity (e.g. the optimal number of repair men in a repair shop) subject to a budget or system availability constraint. However, there is also some research on the topic of flexible repair capacity. De Haas (1995) presents an overview of this line of research. Different forms of flexible repair capacity exist, such as subcontracting repair work, hiring or leasing

extra repair men, overtime policies, and variable working day policies. De Haas uses the term 'Flexible Manpower Planning' (FMP) and defines it as follows:

"The instrument that directs all measures to accomplish variations in the manpower on the short term. The instrument is under the authority of the manager and can be applied with a short lead time".

In his thesis De Haas puts forward two research questions with regard to the use of FMP in repairable item inventory systems:

- 1) How effective is the use of a FMP in repairable item systems, measured in terms of a contribution to a service level?
- 2) How can decisions regarding initial stock and FMP be embedded in a framework for the control of repairable item systems?

With regard to the first research question De Haas concludes that the characteristics of the repairable item inventory system itself (e.g. demand rates, repair times) are of more importance to the effectiveness of the FMP policy than the characteristics of the policy itself. He concludes that FMP policies can contribute significantly to an increase in performance of the system when manpower is tight and high target service levels are set. However, he also emphasizes the fact that under the same conditions FMP policies can have a negative impact on the service performance. This is especially true for repairable item inventory systems with relatively high demand, short repair times, and rather rigid FMP policies (i.e. long lead times). With regard to the second research question he develops a framework for control that consists of three levels. At the highest level, the target coordination level, manpower and system stock are roughly balanced. For each stock location and repair department in the system, targets are deduced from a management goal with the aid of a mathematical model. At the middle level, the structural control level, the means to achieve the specified target are evaluated on cost and practical motives for each repair department individually. At this level the acquisition and control of stocks, the employment of repair men, the selection of priority scheduling rules and FMP rules are determined. At the lowest level, the operational control level, repair work orders are dispatched and their progress through the repair shop is monitored.

2.6 Positioning and discussion

The research presented in this thesis is closely related to the METRIC modelling technique as described in Sections 2.2 and 2.3. The models presented in Chapters 4, 5, and 6 of this thesis also

Literature review

assume ample repair capacity or inventory supply capacity at the highest echelon in the system under consideration. The emphasis, however, is on modelling repair or supply flexibility in these systems. Instead of assuming some fixed policy for satisfying demand for service parts at the lowest echelon in the distribution system, we evaluate different policies with different levels of flexibility with respect to efficiency and effectiveness. This approach is related to some of the work presented in Section 2.5 where different repair policies (scheduling rules in the repair shop) and allocation policies (where to send repaired items?) were evaluated.

Chapter 3

Service Part Supply System: a framework for control

3.1 Introduction

In this chapter we present a framework for the control of repair and supply flexibility in complex distribution and repair networks for service parts. In order to ensure an effective and efficient control of service parts in such networks, we present a framework that captures the most important features of flexibility for service parts. In the framework, called the *Service Part Supply System* (SPSS), we identify several flexibility options that can be applied to increase the performance of the system (see Verrijdt, 1995). For every flexibility option an economic trade-off has to be made that balances the cost of applying flexibility (e.g. cost of using fast direct deliveries) against the increased service performance of the system as a result of applying this flexibility (e.g. shorter response times to customers and therefore lower penalty cost).

In Section 3.2 we present the framework and identify six general flexibility options that can be applied. We decompose the general network structure into two substructures: the repair structure and the distribution structure. For both substructures we discuss three flexibility options that can be applied. In Section 3.3 we discuss the economic trade-offs that are associated with each of the identified flexibility options. In Section 3.4 we relate our framework for the control of service parts to the framework for production control by Bertrand *et al.* (1990). In Section 3.5 we present five case studies and discuss the results with respect to our framework for control. Finally, in Section 3.6 we draw some final conclusions.

3.2 The Service Part Supply System

The Service Part Supply System is a logistic framework for the control of service parts. The framework is presented for the case of repairable service parts but can be used for consumable service parts as well. The case of consumable service parts is discussed at the end of this section. The SPSS describes the operational processes that service parts go through in a repair and distribution network (see figure 3.1). The installed base of technical systems at customer sites is supported by an extensive distribution network of inventory locations that stock service parts. The distribution system depicted in figure 3.1 represents a three-echelon network: local warehouses (located nearest to the customer), national warehouses (responsible for replenishing the local

warehouses), and the central warehouse (responsible for replenishing the national warehouses). In case of repairable service parts, as we assume here, the central warehouse inventory is replenished

- 1: **Return Flow Flexibility**
- 2: **Work Order Release Flexibility**
- 3: **Repair Shop Flexibility**
- 4: **Allocation Flexibility**
- 5: **Pooling Flexibility**
- 6: **Direct Shipment Flexibility**

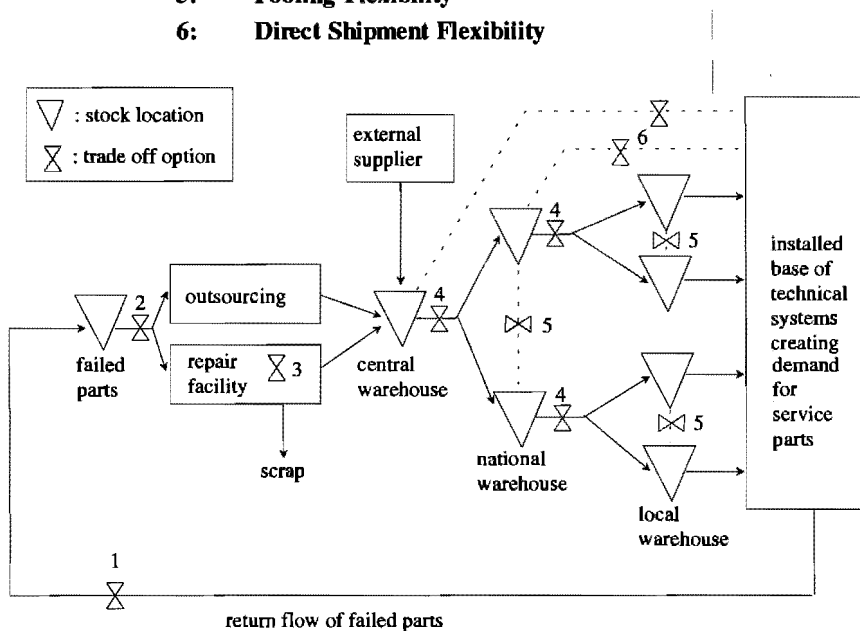


Figure 3.1: Service Part Supply System

by repaired service parts that come from internal or external repair centers. Sometimes defective parts cannot be repaired anymore (due to e.g. wear-out) and are scrapped. In that case new service parts have to be bought at external suppliers. In this way the total system stock can be kept at a fixed level. The number of echelons in for example a European distribution system varies depending on the market situation. A contractually agreed short response time of e.g. 4 hours requires that the inventory locations are located nearby the customer. This necessitates the use of a multi-echelon distribution system with national warehouses and sometimes regional warehouses. However, when the contractually agreed response time is 24 hours, it is possible to use a single-echelon distribution system consisting of one central European warehouse from where the European market is serviced. Note that next to the market requirements (in terms of response times) the trade-off between

inventory holding cost and transportation cost plays an important role as well. Even when the market requirements allow the use of one central European warehouse, it can be cost-effective to use a multi-layered distribution system because of economies of scale in transportation.

Next to the distribution subsystem (consisting of a network structure of inventory locations) we have the repair subsystem. This repair subsystem of the SPSS consists of the return flow of failed service parts from the field to the repair centers, the inventory location where the failed parts are temporarily stocked, and the repair process itself. The repair facility can be a complex structure of mutually dependent repair departments. Complex products (e.g. aircraft engines) generally have a hierarchical structure. The end-item consists of modules which in their turn consist of parts. For each level in the product hierarchy (product-level, module-level, part-level) separate repair centers may exist. The interdependency between these centers is obvious: when repairing at product-level (or module-level) serviceable modules (or parts) are necessary to replace failed modules (or parts). The performance of the repair process in total is then dependent on a series of interrelated repair activities (see also De Haas and Verrijdt, 1996).

In both the distribution subsystem and the repair subsystem we can identify a number of flexibility options that can be applied to increase the performance of the system in terms of cost and service. In figure 3.1 these flexibility options are numbered 1 to 6. We distinguish between flexibility options in the repair subsystem (options 1, 2, and 3) and flexibility options in the distribution subsystem (options 4,5, and 6). Flexibility applied in the repair subsystem we term *repair flexibility* since it concerns different aspects of the repair process of failed parts. Flexibility applied in the distribution subsystem we term *supply flexibility* since it concerns different aspects of the supply process of serviceable parts. We now explain and discuss these six flexibility options in more detail.

Repair flexibility:

- 1) *Return Flow Flexibility*
- 2) *Work Order Release Flexibility*
- 3) *Repair Shop Flexibility*

The return flow of failed parts from the field to the repair center can be controlled in different ways. Failed parts can be shipped directly to the repair center on a one-by-one basis. Another possibility is to batch failed parts and send the batches directly to the repair center. Instead of shipping failed parts (one-by-one or in batches) directly to the repair center, it is also possible to use the existing distribution system to collect and return failed parts. That is, failed parts are first sent to the nearest local warehouse, from there they are returned to the national warehouse and then to the central warehouse, and finally they are sent to the inventory location for failed parts.

The release of failed parts from the inventory location to the repair process contains three different aspects: (i) what part type to repair, (ii) how many parts to repair, and (iii) where should the repair be carried out. The inventory location contains many different types of parts that are waiting for repair. The decision what type to repair can be based for example on the echelon inventory positions of the different part types. The decision how many parts to repair can be based on the structure and capacity of the repair shop. The decision where to carry out the repair activities implies that a choice has to be made between outsourcing the repair orders or using your own repair facilities. This decision can be based on utilization rates and planned workload of the repair shop.

When the decision is taken to perform the repair activities in your own repair facility, an additional number of flexibility options can be applied. For example, it is possible to use different priority scheduling policies and batching policies to influence the performance of the repair shop. Another option is the use of flexible manpower planning such as overtime policies and variable working days policies (De Haas, 1995).

Supply flexibility:

- 4) *Allocation Flexibility*
- 5) *Pooling Flexibility*
- 6) *Direct Shipment Flexibility*

The allocation of inventory in the supply chain is an important decision variable. We distinguish three levels of Allocation Flexibility. At the highest (strategic) level the decision has to be made what parts to stock in the various inventory locations in the supply chain. This so-called 0-1 decision is very important for service parts logistics because of the large size of the assortment in combination with the high-price low-demand characteristics. At the middle (tactical) level the stock levels for the parts that are taken into the assortment (1-decision) must be determined for every inventory location in the supply chain. Finally, at the lowest (operational) level an allocation policy must be specified that allocates available inventory in case of shortages.

When a demand arrives at a local warehouse that is out of stock, pooling flexibility prescribes to check neighboring local warehouses at the same echelon for excess stock. Viewing inventories of local warehouses in the same region as one inventory pool might be very rewarding. Lateral transshipments between local warehouses in the same pool can reduce the waiting time for customers significantly. It is also possible to consider pooling flexibility at a higher echelon in the supply chain (e.g. sharing of inventory between national warehouses). Finally, it is possible to define different conditions under which pooling flexibility is allowed. Instead of applying only pooling flexibility when a stock-out situation at an inventory location occurs (i.e. trigger level for pooling is zero), it

Service Part Supply System

is also possible to define positive trigger levels that allow the use of emergency lateral transshipments.

Direct shipment flexibility allows the use of emergency shipments of service parts from an inventory location at a higher level in the supply chain (i.e. national or central level). Instead of using the normal replenishment channels (with relatively long lead times) it is possible to use fast direct deliveries when necessary. When direct shipment flexibility is allowed in stock-out situations at local warehouse, it can reduce the waiting time of customers significantly. The main characteristic of direct shipment flexibility is that fast deliveries of service parts are allowed that skip one or more echelons in the distribution system.

Consumables

We end this section with a brief discussion on the difference between consumables and repairables with respect to the SPSS framework. The framework presented in this section describes the processes that repairable service parts follow. The subsequent flexibility options discussion is also based on the assumption of repairable service parts. However, the framework is applicable as well for consumable service parts. Instead of repairing failed service parts, the failed parts are scrapped and new parts are procured at a vendor. The repair flexibility options (return flow flexibility, work order release flexibility, and repair shop flexibility) do not exist in such a situation. The main advantage for consumable parts is that the system inventory can be reduced more easily by stopping the procurement process of new parts. However, due to long term contracts and low demand frequencies of service parts in general, the reduction of systemwide inventory is even for consumables very hard.

3.3 Economic trade-offs

The application of flexibility as discussed in the previous section influences the performance of an SPSS in terms of cost and service. For each of the six flexibility options an economic trade-off must be made to balance the additional cost of using flexibility against the increased service performance as a result of using flexibility. In this section we discuss for each of the six flexibility options the economic trade-off that has to be taken into account for the implementation decision of the specific flexibility option. The benefits of applying flexibility (in terms of e.g. shorter response times or inventory reduction) must outweigh the extra effort that has to be paid for using flexibility (in terms of e.g. higher transportation cost or increased nervousness in the planning).

1) *Return Flow Flexibility*

The choice between returning failed items at the customer's sites on a one-by-one basis or in larger batches to the central repair shop is based on the trade-off between repair-cycle-time length and transportation costs. The repair-cycle-time is defined as the time between failure of a service part and the completion of repair of that particular part. A minimal repair-cycle-time, which is the case when returning parts on a one-by-one basis, minimizes the total system stock. The transportation cost on the other hand is higher since every part is shipped individually to the repair shop immediately after failure. When returning failed items in batches the reverse situation is true. Transportation cost is lower but the repair-cycle-time is longer since failed items are not immediately returned to the repair shop.

Another decision with respect to the control of the return flow of failed parts is the decision on how to organize the return flows. Should failed parts be sent directly to the central repair shop (as depicted in figure 3.1) or should the existing distribution structure be used for collecting and returning failed items. The trade-off is again between transportation cost and repair-cycle-time. Using direct shipments minimizes the return time of failed parts but is very costly. Using the existing distribution network lengthens the return time but is cheaper.

2) *Work Order Release Flexibility*

The trade-off that has to be taken into consideration for using a work order release mechanism for repair orders is between the benefits of controlling the size and the mix of the input flow into the repair shop (in terms of throughput times and utilization rates in the repair shop) and the effort that must be made to implement it (in terms of planning and information requirements). An uncontrolled flow of orders into the repair shop will lead to a lengthening of the repair cycle time and therefore to an increase in system inventory. Controlling the flow of repair orders, for example by subcontracting repair work, will increase the performance of the repair shop. In practice, some kind of release mechanism is always used for controlling the input flow. The question therefore usually is how sophisticated should the release mechanism be. Should it incorporate all three aspects (what parts, how many, and where to repair) or only a subset of these aspects.

3) *Repair Shop Flexibility*

The trade-off that is associated with the various kinds of repair shop flexibility options is similar to the previous one. Parts must be repaired, no matter what, but how sophisticated should the shop floor control mechanism be? Is it worth implementing complex priority scheduling rules or batching rules for the repair shop under consideration? The structure of the repair shop determines to a large

Service Part Supply System

extent the effectiveness and efficiency of the different flexibility options. De Haas (1995) for example shows that the effectiveness of using variable working day policies in repair shops primarily depends on the characteristics of the system under consideration. The characteristics of the various policies themselves is less important.

4) *Allocation Flexibility*

The strategic 0-1 decision for determining the assortment of service parts for every inventory location is based on the trade-off between the cost of stocking a service part and the cost of not stocking the service part. The cost of stocking a particular part depends on the price of the item, the interest rate, the space that is needed to store the item, and the risk of obsolescence. Especially this last factor is very important since service parts show in general a very low demand rate. The cost of not stocking a particular part depends on the effort that has to be made for procuring the part when it is needed. It is possible that the part can be procured by using an emergency shipment from another inventory location. It is also possible that the part has to be manufactured in which case the cost can be very high, especially when the part is no longer incorporated in the current production planning and a new production line has to be set up. The cost of not stocking a part is also affected by the consequences of not having that part immediately available when needed. If an airplane is grounded because of a missing service part, the financial consequences can be very dramatic. If a copier is out of order because of a missing service part, the consequences are less dramatic.

The tactical decision on determining the stock levels for service parts that are taken into the assortment is based on the trade-off between inventory holding cost and penalty cost for not having the part available when needed. This trade-off is to some extent similar to the previous one. For most service parts, however, the stock level for parts that are stocked is one because of the low demand rate. Service parts that have a high demand rate are usually less critical and less expensive. Determining the stock levels for these fast moving items is therefore less critical.

The operational decision on allocating service parts in case of shortages is usually based on circumstantial arguments. The trade-off is again between the level of sophistication of the allocation rule (and its associated planning and information requirements) and its contribution to the service performance of the system as a whole.

5) *Pooling Flexibility*

The application of pooling flexibility for local warehouses in the same region is based on the trade-off between the extra cost of using emergency lateral transshipments and the increase in service performance. The cost of emergency lateral transshipments for individual service parts between local

warehouses in the same geographical area is usually based on fixed tariffs. Special courier services like DHL, TNT, and UPS have standard tariffs that are based on weight and distance. The benefit of pooling flexibility lies in the fact that the response time towards the customer is reduced. If a customer is waiting for a specific service part that is not available at the nearest local warehouse, a lateral transshipment from another local warehouse in the same region or country is the fastest way to satisfy this backorder. The cost of the lateral transshipment has to be balanced against the waiting cost for the customer. Another important element of this trade-off is that the sourcing local warehouse in a pooling group temporarily reduces its own service ability by acting as a supplier to another local warehouse.

6) *Direct Shipment Flexibility*

Application of direct shipment flexibility is also based on the trade-off between the cost of direct deliveries and reduction in response time towards the customer. The cost of direct deliveries are usually higher than the cost of lateral shipments since the distance over which the service parts are transported are in general significantly longer. Therefore the trade-off considerations are identical but the outcome of the trade-off can differ. Service parts that are eligible for pooling flexibility are not necessarily eligible for direct shipment flexibility.

An overview of the relevant cost factors that are involved with the economic trade-off of each flexibility option is presented in table 3.1. Next to the four operational cost factors we identified in Chapter 1 (inventory holding cost, normal transportation cost, emergency transportation cost, and penalty cost for shortages), we mention the cost of IT (Information Technology). Implementation of any kind of flexibility in a Service Part Supply System will require extra investment in IT. The extent of the IT-investments depends on the type of flexibility that is implemented. IT-investments are in general long term decisions with a high economic risk. The investment cost for IT is therefore not an operational cost factor. Implementation of operational flexibility should be based on an economic trade-off of operational cost factors. However, this implementation of operational flexibility is dependent on the structural decision to invest in IT.

Table 3.1 shows how the different types of flexibility can be applied in order to reduce for example the systemwide inventory and therefore also inventory holding cost (denoted by ↓). Return flow flexibility can be applied in order to reduce repair cycle times. This implies that the number of return shipments increases and the size of the return shipments decreases. Consequently the transportation cost (either normal or emergency, depending on the type of transportation that is used) will increase. Both work order release flexibility and repair shop flexibility improves the planning and scheduling of repair activities. As a result the inventory buffer can be decreased since the repair activities are better tuned to the distribution activities. The cost of these two types of flexibility is

mainly caused by investment in sophisticated information and planning tools. Finally, the supply flexibilities (allocation, pooling, and direct shipment) improve the availability of stock in the supply system. Stock is no longer allocated to one specific warehouse but can be used by other warehouses as well. Consequently, less stock is needed in the system as a whole. Moreover, the response times toward customers is minimized and therefore also the penalty cost decreases. The cost of implementing these flexibilities is primarily caused by the cost of using emergency shipments throughout the supply system and the cost of using information systems that can track and trace inventory throughout the supply system.

| FLEXIBILITY OPTION | COST FACTORS | | | | |
|--------------------|-------------------|------------------|---------------------|------------------|----|
| | inventory holding | normal transport | emergency transport | shortage penalty | IT |
| Return flow | ↓ | ↑ | ↑ | | ↑ |
| Work order release | ↓ | | | | ↑ |
| Repair shop | ↓ | | | | ↑ |
| Allocation | ↓ | | ↑ | ↓ | ↑ |
| Pooling | ↓ | | ↑ | ↓ | ↑ |
| Direct shipment | ↓ | | ↑ | ↓ | ↑ |

Table 3.1: Overview of the relevant cost factors with respect to the flexibility trade-offs

3.4 Embedding the SPSS framework

In this section we relate our framework for the repair and distribution control of service parts to the framework for production control as described in Bertrand *et al.* (1990). Although the two environments that are considered differ significantly from each other, we adopt the main principles that are identified for a production environment and translate them to our repair and distribution environment for service parts. A complete and absolute correspondence between the two frameworks is impossible, but it is possible to use the general ideas that are being applied in the framework by Bertrand *et al.* We now first describe the main principles applied in the framework for production control of Bertrand *et al.* Next we show how some of these principles can be translated to our Service Part Supply System.

Bertrand *et al.* present a framework for the design of a production control structure that is based on three important characteristics.

- 1) *Goods Flow Control vs. Production Unit Control*
- 2) *Aggregate Production Planning versus Material Coordination*
- 3) *Relationship between Production and Sales*

First, they make a distinction between *Goods Flow Control* and *Production Unit Control*. Production Units (PU) are referred to as more or less independent production departments which on the short term are self-contained with respect to the use of their resources, and which are responsible for the production of a specific set of products from a specific set of materials and components. Goods flow is referred to as the flow of physical work orders between Production Units. Production Unit Control is concerned with the realization of an agreed performance (in terms of e.g. throughput time of work orders) for a PU, given specific environmental conditions. Goods Flow Control is concerned with the problem of how to realize for each PU the agreed environmental conditions, and to realize the overall production control objectives (in terms of e.g. delivery performance and flexibility to the market) at the same time. From the Goods Flow Control point of view, the PU's are considered black boxes that perform according to specified norms. A graphical representation of the relation between Production Units and Goods Flow Control is presented in figure 3.2.

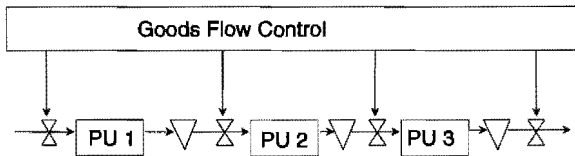


Figure 3.2: *Relation between Production Units and Goods Flow Control*

Second, at Goods Flow Control level a distinction is made between *Aggregate Production Planning* and *Material Coordination*. Aggregate Production Planning is mainly concerned with the coordination and control of aggregate parameters (e.g. bottleneck capacity, production budgets, inventory budgets) on the long term. Material Coordination is concerned with the allocation of these aggregate parameters to the production of individual product items on the short term.

Third, the relationship between Production and Sales is included in the control framework. The coordination of these two entities is part of the Goods Flow Control. Usually quotations are formulated between Sales and Production in terms of quantity of sales/production per family, delivery conditions, reliability of sales forecasts etc. A structured coordination mechanism between Production and Sales is a vital element of the production control framework.

The SPSS framework

In our SPSS framework we can also make a distinction between the control of more or less independent "Production Units" and the control of goods flows between these units. We refer to these units as *Activity Units* (AU) instead of *Production Units* since the processes under consideration are not necessarily manufacturing processes. The Activity Units in an SPSS are the individual inventory locations and the repair centers. Activity Unit Control is concerned with the realization of an agreed performance in terms of service and cost. The coordination of the processes between the Activity Units is performed by the *Supply Chain Control* function. This control function is equivalent to the Goods Flow Control function in the production control framework. A graphical representation of the relation between Activity Units and Supply Chain Control is presented in figure 3.3.

- AU 1 : inventory location for failed parts**
- AU 2 : repair center and central warehouse**
- AU 3 : national warehouses**
- AU 4 : local warehouses**

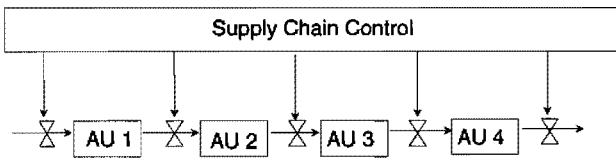


Figure 3.3: Relation between Activity Units and Supply Chain Control

Supply Chain Control is responsible for the return flow coordination of failed parts to the inventory location for failed parts (AU 1), the work order release function for the repair shop (and central warehouse, AU 2), the allocation policy for the parts shipped to the national warehouses (AU 3), the allocation policy for parts shipped to the local warehouses (AU 4), and the supply of service parts to the customer in the field.

At the Supply Chain Control level a distinction can be made between an aggregate planning level and a detailed planning level (analogous to the Aggregate Production Planning function and Material Coordination function in the production framework). At the aggregate planning level long term decisions are taken with respect to the design of the SPSS. These decisions include for example the capacity determination of the Activity Units (repair capacity, warehouse capacity) and the assortment composition at each inventory location in the supply chain. The decisions taken at the aggregate level specify the environmental conditions at which the decisions at the detailed planning level are

taken. The decisions taken at the detailed planning level can make use of some of the flexibility options that are identified in the SPSS framework. These decisions are based on trade-offs between cost and service performance and prescribe for example how to allocate parts in case of shortages. They also prescribe how to act in case of a stock-out situation at a local warehouse, e.g. apply lateral transshipments or direct deliveries.

The use of maintenance strategies and concepts (Gits, 1984) supports the coordination between Supply and Sales of service parts in our framework (analogous to the coordination between Production and Sales in the production framework). Maintenance strategies and service contracts influence the demand for service parts and are therefore vital for both Supply and Sales. Concepts like preventive or periodic maintenance and leasing of products can be applied to increase demand information on service parts. This kind of information on installed base characteristics supports the inventory decisions that have to be taken for the service parts.

3.5 Case studies

In this section we discuss five case studies that were carried out as graduation projects in the field of service parts logistics. All five case studies are briefly described and the main results are summarized. The projects and its outcomes are reviewed with respect to the different kind of flexibilities identified in the Service Part Supply System. At the end of this section we summarize our general findings with respect to these case studies.

3.5.1 Intergraph case

Description

The first case was carried out at Intergraph and is described in detail in Lamers (1994). Intergraph Corporation is an American company specialized in graphical computer hard- and software for industrial and professional applications. The project was carried out at the Field Services department in Nijmegen, responsible for the after sales service and service part distribution in Europe. The service part supply system of Intergraph consists of three echelons: one central warehouse in Nijmegen, national warehouses in 24 European countries, and several regional warehouses associated to each national warehouse. Field engineers pick up service parts at the regional warehouses and visit customers to solve machine problems. The aim of the project was to gain insight into the parameters that determine the service performance of the service part supply system and the associated costs. The current logistic concept was evaluated and alternatives were proposed. The attention was focused on the Nijmegen - UK branch of the service part supply system.

Modeling technique

The way in which the service performance of the supply system is measured is very important when evaluating different logistic concepts. In this case study we looked at two different service measures. First, the traditional *fill rate* (FR) measure was calculated using an adapted version of the METRIC model. This measure indicates the fraction of demand per part that can be met from stock on hand. A hierarchical approach was used to calculate the fill rates of the stocking locations at the different echelons. The fill rate of the UK as a whole, for example, was calculated by aggregating all regional warehouses in the UK into one warehouse and comparing the resulting fill rate with the average fill rate of the individual regional warehouses. Second, we calculated the fraction of jobs, or work orders, that service engineers can execute without delay due to missing service parts. We term this measure the *work order completion rate* (WCR). The execution of a work order may require more than one part. The probability that a work order can be executed without delay is approximated by multiplying the probabilities that each of the individual service parts needed for the work order is available from stock on hand (i.e. the individual fill rates). This measure is an approximation since it is assumed that the demand for all service parts is independent of each other. The relation between the work order completion rate and the fill rates can be expressed as follows:

$$WCR = \sum_{i=1}^{\infty} \alpha_i FR^i$$

where α_i represents the fraction of the work orders that consists of i parts.

Although the complete assortment of active parts (approximately 1300 parts) was stocked in the UK, the work order completion rate was estimated to be only 15 %. This was caused by the fact that most repair work orders required (slow-moving) parts that were stocked centrally in the UK and therefore overnight supply was needed. Since overnight supply was possible as well from the central European warehouse in Nijmegen, a trade-off model was constructed to balance the cost of inventory against the cost of transportation for different stocking policies.

Conclusions

In general we concluded that reducing replenishment lead times for spare parts put on stock in the UK had a negligible impact on the service performance. Parts put on stock in the UK already have a high fill rate. The relatively long periods of time between two consecutive demands for a service part (in general more than ten weeks) in combination with relatively short replenishment lead times (approximately one week) does not leave much room for increasing the service performance even

further. The most important factor determining the service performance of the service part supply system is the decision *what* parts to put on stock in the regional warehouses. For example, suppose the fill rate for a regional warehouse equals 30%. The work order completion rate can then be calculated to be approximately 16%. Increasing the availability of the parts put on stock in the regional warehouse from e.g. 90% to 95% will result in an fill rate of 32% and a work order completion rate of 17%. This illustrates the limited effect of improving the availability of parts already put on stock even further on the performance indicator WCR.

Evaluation of the current logistic concept revealed that the information system that was used for the control of service parts showed some serious deficiencies. The stock norms for the various stockpoints were not traceable and demand information (usage and consumption data) was distorted. Finally, the model that was used to compare different stocking policies, taking into account inventory cost and transportation cost, showed that it was optimal to allocate the (approximately) 150 most active parts in the UK and reallocate the remainder of the assortment to Nijmegen. This analysis was done for different transportation tariffs for overnight supply from Nijmegen to the UK. Since prices of service parts were expected to decrease in the future the same analysis was also done for lower average prices of service parts. Both scenarios showed that it was still rewarding to reallocate slow-moving service parts to Nijmegen. The active parts that are held in stock in the UK should be allocated well to the different regional warehouses such that the service performance is maximized.

Relation with the SPSS

This case study addresses the use of supply flexibility in the service part supply system of Intergraph. Pooling flexibility was already applied since regional warehouses in the UK can supply each other with service parts in case of shortages. Direct shipment flexibility was also applied by sourcing parts from the national warehouse or from the European warehouse in case of shortages. The main focus in this project was on the decision how to allocate stock in the supply chain. Evaluating different stocking policies under different scenarios revealed that considerable cost savings can be obtained by reallocating stock in the central warehouse in the UK: centralize slow-moving parts and decentralize fast-moving parts. Applying allocation flexibility showed that the use of a central national warehouse in the UK is unnecessary. Given the transportation times and cost within the UK and between Nijmegen and the UK, the European warehouse in Nijmegen can actually take over the place of the national warehouse in the UK.

3.5.2 ASM Lithography case

Description

The second case was carried out at ASM Lithography and is described in detail in Kanter (1994). ASML is a high tech company developing, assembling, selling, and servicing wafer steppers, which are used for the production of integrated circuits. ASML is one of the three major producers of wafer steppers in the world and its main market is the USA. ASML Service is responsible for the preventive and corrective maintenance activities. Corrective maintenance is performed according to the repair-by-replacement principle for which service parts are needed. ASML Logistics is responsible for the world wide supply of the required service parts. The service part supply system of ASML consists of three echelons: a centralized worldwide stock in Veldhoven in the Netherlands, a continental stock for the USA in Tempe Arizona, and local warehouses in Europe, Asia, and the USA located near customers. The aim of the project was to develop a model for stock decisions for the different categories of service parts at the various locations in the supply system, taking into account customer requirements and expectations.

Modeling technique

The concept of decoupling points was introduced to facilitate the decision on how to compose the assortment of service parts at the different levels in the supply system. Several characteristics play a role in this decision, such as cost price of a part, demand frequency, risk of obsolescence, required delivery time and reliability, and replenishment time. The market in this line of business requires very short response times since the production processes of many customers are dependent on the functioning of the wafer steppers. Consequently, the down time cost of these systems is very high. In order to support the allocation decision quantitatively, the METRIC modeling technique called *marginal analysis* was applied to determine the assortment of service parts at the local warehouses. Service parts with the highest contribution in decreasing backorders per invested guilder are added to the local inventory, until a desired *operational availability* (or up-time percentage) of the wafer steppers is realized. This operational availability is calculated as the product of two separate availabilities: *maintenance availability* (determined by factors such as number of maintenance engineers, available test equipment, and overall maintenance policy) and *supply availability* (determined by the stocking policy for service parts needed for maintenance). This supply availability is used in the model and is determined as follows (Sherbrooke, 1992b):

$$\text{supply availability} = \prod_{i=1}^I \left(1 - \frac{EBO(S_i)}{NZ_i} \right)^{Z_i} * 100 \%$$

with I = number of part types
 Z_i = number of occurrences of part i on a wafer stepper
 N = number of wafer steppers
 $EBO(S_i)$ = expected backorders of part type i for stock level S_i
 $(0 \leq EBO(S_i) \leq N Z_i)$

The term $(1 - EBO(S_i) / N Z_i)$ represents the availability of service part i . This term to the power Z_i represent the availability of a wafer stepper due to part i (the number of occurrences of part i on a wafer stepper is equal to Z_i). Finally, multiplying over all service parts ($i=1..I$) gives the general expression for the availability of a wafer stepper as a result of the stocking policy for service parts. The same modeling technique was used to determine the stock composition at the continental warehouse in the USA. The fill rate of the continental warehouse should be 100% since this is assumed when determining the local inventories. However, because of the possibility of emergency shipments, a less higher service degree is acceptable (e.g. 98%). Applying this modeling technique results in a stock decrease at the local and continental level in the supply system. To compensate this effect extra safety stock is needed at the worldwide inventory location in Veldhoven.

Conclusions

The absence of a planning tool to support stocking decisions for service parts at the local warehouses (and the USA continental warehouse) was tackled by introducing the concept of decoupling points that gives an indication on where to stock service parts in the supply chain. The technique of marginal analysis was used to determine the optimal stock levels that are needed to realize a predetermined operational availability of wafer steppers. This relation between stock decisions and service performance (i.e. operational availability) was very important. Application of the model was expected to result in significant reduction of service part inventories. Part of these savings must be used to increase the back-up performance of the worldwide inventory in Veldhoven.

Regarding the implementation of the proposed policy, internal (within ASML) and external (customers) acceptance of the marginal allocation tool is essential. Although this is not necessarily true, a high fill rate at the local warehouses is often believed to be essential for realizing high operational availability. Customers also can demand a high fill rate for service parts. In these situations it is also possible to use the marginal analysis tool for determining stock levels when aiming for a target fill rate.

Relation with the SPSS

In this case study we examined the possibilities to support stocking decisions for different levels in

Service Part Supply System

the service part supply system of ASM Lithography. This corresponds to the allocation flexibility option in the SPSS framework. We addressed the strategic issue of where to allocate service parts by introducing decoupling points in the supply chain. The tactical issue of determining the optimal stock levels for parts was solved by applying the METRIC technique of marginal analysis. An important notion that emerged from this case study was the fact that different service objectives can exist for service part supply systems. Depending on internal (within the company) and external (customers) circumstances, different stocking decisions may have to be taken.

3.5.3 Siemens case

Description

The third case was carried out at Siemens Nederland and is described in detail in Sengers (1995). Siemens is a German company specialized in electronics for the industrial and consumer market. In the Netherlands Siemens has seven business units each with their own specific product-market combination. Each business unit is responsible for their own service activities. For reasons of efficiency the service part inventory of the seven business units is centralized in Zoetermeer. The department Logistics Material Management Service (abbreviated Log MM Service) is responsible for ordering parts at the vendors (usually Siemens Germany), handling and warehousing of these parts upon arrival at the central warehouse in Zoetermeer, and disposition of service parts to engineers of the seven service organizations. This disposition concerns normal replenishment of car stocks and emergency replenishment in case of stock out situations. The service part supply system of Siemens in the Netherlands consists of a central warehouse in Zoetermeer which supplies 80 service technicians with service parts. These technicians visit customers to solve machine problems. The central inventory consists of kits (for diagnostic use) and parts. The main supplier of the central warehouse is Siemens Germany. The seven service organizations are responsible for introducing new service parts in the central inventory, indicating the initial stock level for these parts (called ORB), and removing service parts from the assortment. They are also responsible for the car stocks of the service technicians. Log MM Service is responsible for guarding and adjusting the stock levels using the demand data of the service parts.

The aim of the project was to develop a method for determining optimal stock levels for the central service part inventory, taking into account the trade off between service performance and associated costs. Therefore, a suitable definition and measurement of the service performance was needed first. The project was focused on the relation between Log MM Service and the service organization ITS (Information Technology Service).

Modeling technique

To analyze the cost of the central warehouse in Zoetermeer five categories of cost were distinguished and calculated: 1) overhead cost for the department Log MM Service (26%), 2) overhead cost for the department Physical Distribution (19%), 3) depreciation cost for obsolete stock (33%), 4) transportation cost for distributing parts from the central warehouse to the service engineers (13%), and 5) interest cost for the physical stock in the warehouse (10%). The service performance was measured in three ways: 1) number of emergency orders for parts needed by the engineer at the customers' site, 2) analyzing data from the information system that records malfunction at customer locations, and 3) measuring the efficiency and effectiveness of the tasks performed by Log MM Service (in terms of e.g. replenishment times for normal and emergency orders or adjusting stock levels).

The model that was used for the calculation of the optimal stock levels in the central warehouse is based on a model by Muckstadt and Thomas (1980) that incorporates emergency shipments in case of stock outs. The allocation of inventory to the central warehouse is based on a marginal analysis technique that balances the realized fill rate performance and the investment in inventory. The demand data used in the model reflected the actual consumption of parts and not the usage of parts (e.g. for diagnostic purposes). As a result two opposite effects occur. On the one hand the calculated fill rate is an upper bound for the actual fill rate since the demand for parts is higher than estimated. On the other hand the fill rate increases because parts that return unused have triggered a replenishment and therefore the inventory increases.

Conclusions

The cost analysis showed that the highest cost savings could be obtained by reducing inventory and therefore reducing depreciation cost and interest cost. Furthermore it was shown that transportation cost could be reduced significantly by allowing non-standard (with respect to size) parts to be distributed by in-night service instead of courier services. The measurement of the service performance showed that analysis of emergency orders can be used to adjust the stock levels in the central warehouse and therefore increase the service performance. The data retrieved from the information system that records malfunctions proved to be unreliable.

Application of the stock model resulted in optimal stock levels for parts ranked to price and demand data. The optimal stock levels range from zero (for expensive parts with a low demand) to six (for cheap parts with a high demand). This theoretical result was adjusted in consultation with the service department ITS in order to realize high service performance for all parts.

Relation with the SPSS

In this case study we investigated the calculation of optimal stock levels for the central warehouse in the service parts supply system of Siemens Nederland. This decision is part of the allocation flexibility as identified in the SPSS framework. We did not address the strategic issue of determining the composition of the central stock (0-1-decision), since this decision is made by the service organizations who have more and better information. The tactical decision of determining the optimal stock levels for parts is based on a trade-off between demand frequency and cost price.

3.5.4 Philip Morris case

Description

Case number four was carried out at Philip Morris Holland B.V. in Bergen op Zoom and is described in detail in Heijman (1996). Philip Morris produces approximately 76 billion cigarettes every year at its production plant in Bergen op Zoom. Ninety percent is destined for export to other countries in Europe, mainly Italy and France. The production process of cigarettes exists of two main phases that are executed in different departments. In the Primary Department the tobacco is prepared by putting together different raw materials. In the Secondary Department the cigarettes are produced (from the prepared tobacco) and packed. The project was carried out at the Maintenance Support Primary (MSP) Department which is responsible for all maintenance activities carried out in the Primary Department. The aim of the project was to support and to improve the logistic control of service parts needed for maintenance activities in the Primary Department. The service part supply system for Philip Morris consists of an international warehouse in Switzerland that acts as supplier for all European production plants, three internal stock locations at the production plant in Bergen op Zoom, and a number of unregistered "grey" stock locations at several places in the factory. The goal of the project was to design a decision support system that helps the MSP Department in making the right stocking decisions for service parts. Furthermore, performance indicators were to be developed that measure, evaluate and improve the logistic performance of the service part control processes.

Modeling technique

To support the 0-1 stock decision for service parts a model was developed in four steps. First, decision criteria that play an important role in whether or not to stock parts were established in consultation with the Primary production managers. The following three criteria were defined: 1) *criticality* (how critical is a service part for the production process?), 2) *availability* (what alternatives exist to replace or substitute a service part?) and 3) *cost* (trade-off between stocking a

service part and using direct delivery shipments from the supplier). Second, the criteria were ranked in consultation with production managers. Third, each criteria was defined in such a way that it could be used as decision variable in the model. Criticality was defined as a weighted sum of the length of unplanned down time, decline in the quality of the tobacco, consequences for personnel and environment, and secondary damage at other equipment in the department. Availability was defined as a weighted sum of internal and external sources that can be applied to procure a specific service part in case of a stock-out situation. Examples of internal sources are cannibalization of other equipment, using substitute parts or back-up systems, and internal repair possibilities. Examples of external sources are emergency shipments from other Philip Morris plants or directly from the suppliers. Finally, the cost criteria was based on a trade-off between on the one hand inventory cost (I-decision) and on the other hand direct delivery cost and cost of overtime to meet the production targets in case of production down time (O-decision). The final step in the development of the model was the evaluation step in which a selection of service parts was used to run the model.

Conclusions

The model developed in this project can be used to support the decision whether or not to stock service parts. Application of the model results in a stock recommendation that is based on a sound trade-off between the criticality, the availability, and the replenishment cost of service parts. A sample proof of twenty items was used to test the model. According to the model six items were stocked unnecessarily, representing a stock value of approximately 75.000 Dutch guilders.

The second part of the assignment consisted of the development of performance indicators that could be used to measure, evaluate and improve the logistic flows of service parts within the factory. The main purpose of this exercise was to improve the communication between the different organizational units that were involved with these processes. Several new performance indicators were developed. After a group discussion it was decided by the MSP management to implement the performance indicator that measures the amount of plant-items (service parts that are ordered directly at the supplier and are not stocked at Philip Morris) that were ordered by service engineers, and that were returned (unused) to stock at Philip Morris.

Relation with the SPSS

In this case study we focussed on the allocation decision: whether or not to stock service parts at the production facility itself. This form of allocation flexibility is different from the previously described cases in that it concerns *internal* allocation of service parts. We derived criteria that (in combination) answer the question whether or not to stock service parts. Another distinctive feature

Service Part Supply System

of this case study is the fact that the customer (the Primary Department) and the supplier (Maintenance Support Primary) of service parts are located within the same company. From the perspective of our SPSS-framework it can be noted that the demand processes are therefore more manageable. Although the occurrences of demand for service parts are still highly unpredictable, the source of demand (i.e. machinery at the Primary Department) is completely known.

3.5.5 Digital case

Description

Case number five was carried out at Digital Equipment Corporation (DEC) and is described in detail in Prins (1996). DEC is an American company that develops, manufactures, sells, supports and services computer systems, peripherals and software. The project was carried out at the European Services and Supply Center (ESSC) in Nijmegen which is responsible for the supply of service parts in Europe. The service part supply system of Digital consists of a central European warehouse (i.e. ESSC) that is supplied by numerous suppliers in Europe and the USA, and several warehouses throughout Europe (so-called Service Stocking Points or SSPs) that are supplied directly by the ESSC. Within a country several SSPs can exist that are supplied directly by the ESSC. A more refined network of stocking locations can exist within a country, depending on geographical and market conditions. Subject of research in this project was the flow of service parts from European vendors, via the ESSC, to the SSPs in the different European countries. The goal of this project was to model and evaluate two alternative distribution structures: 1) direct shipments of service parts from European vendors to the SSPs in Europe, and 2) centralization of less critical service part inventories within the Netherlands.

Modeling technique

The direct shipment distribution alternative could only be applied for external suppliers who were capable of delivering service parts directly from stock on hand. Furthermore, the replenishment lead time should be about the same as in the situation where the ESSC acts as central inventory buffer. A model was built to calculate the physical distribution cost of the two distribution configurations. Cost factors that had to be taken into account for the present configuration were transport cost from the suppliers to the ESSC, cost of pipeline inventory and physical inventory on hand at the ESSC, handling cost at the ESSC, transport cost from the ESSC to the national hubs where the incoming shipments are broken into smaller shipments for the various SSPs within that country, handling cost at the national hubs, and transport cost from the hubs to the various SSPs. The cost factors that have to be taken into account for the direct shipment configuration were the transport cost from the suppliers directly to the SPPs and extra ordering cost since consolidation of orders per supplier or

per SSP is not possible anymore. Eight pilot suppliers were selected for a comparison of the two distribution configurations. All parts supplied by a particular supplier (the suppliers menu) were assumed to be eligible for direct shipment to SSPs in six countries in Europe (accounting for 74% of the total demand). The suppliers menu was characterized by the average weight per part, the average price per part, the obsolescence risk, the demand rates, and the suppliers location.

The second part of the assignment concerned the centralization of commodity parts (i.e. less critical service parts that require a response time of one or more days) in the Netherlands. In the present situation these parts were stocked at a local warehouse in Gouda that was replenished by the ESSC in Nijmegen. To investigate the effect of centralization in terms of service performance and physical distribution cost an adapted version of the model by Muckstadt and Thomas (1980) was applied (see also Chapter Two). The adapted model takes into account the fact that sometimes (3 out of 10) parts are returned unused by the service engineer and that service parts are either consumable, repairable locally in Gouda, or repairable centrally in Nijmegen. Two scenarios were evaluated: 1) stock levels at Gouda were set equal to zero and stock levels in Nijmegen remain the same; 2) stock levels at Gouda were set equal to zero and stock levels in Nijmegen were increased with the original Gouda stock levels.

Conclusions

Comparison of the present configuration with the direct shipment alternative showed that four of the eight pilot suppliers were eligible for the direct shipment option. The parts menus of these suppliers were characterized by a low average weight in combination with a high average price. Application of the direct shipment alternative to the other four pilot suppliers would result in a significant cost increase. A quick scan approach was used to determine the indifference line for using the direct shipment option with respect to the average weight and price of a part in a suppliers menu. This was done for six European countries where a supplier could be located. The location of a supplier determines the attractiveness of the direct shipment option as follows (ranked in decreasing order of attractiveness): Germany, the Netherlands, United Kingdom, Belgium, France, and Italy. Finally, a sensitivity analysis showed that these indifference lines for using direct shipment deliveries were to a large extent insensitive to a decrease in transport tariffs from the hubs to the SSPs, a decrease in handling tariffs at the ESSC, and a decrease in direct shipment tariffs to the SSPs. However, the outcome is highly sensitive with respect to a possible increase in purchase price of parts.

The model that was used to evaluate centralization of service part inventories in the Netherlands was validated by comparing the model service performance (expressed in fill rate) with the service performance measured in practice (expressed in work order completion rate). In order to compare

Service Part Supply System

these different service measures, the same transformation was applied as in the Intergraph case. The work order completion rate at the Gouda warehouse calculated from the model was 95.2% whereas the value measured in practice was 93.2%. The service performance at the ESSC (expressed as fraction of order lines delivered completely from stock on hand) calculated from the model was 89% whereas the value measured in practice was 88%. A similar transformation was used as for the work order completion rate service measure. Analysis of the two scenarios described before showed that scenario 1 results in a decrease in average fill rate from 97.4% to 95.1%. The corresponding work order completion rate drops from 95.2% to 92.0%. This decrease in average service performance is solely caused by the large group of slow-moving C-items. The fill rate for these items drops from 95.4% to 84.4%. Scenario 2 results in an increase in average fill rate from 97.4% to 98.6%. The corresponding work order completion rate increases from 95.2% to 97.6%. In order to maintain present service performance for C-items as well, a third scenario was defined: decrease the stock levels for A- and B-items in the Gouda warehouse to zero and reallocate the inventories of C-items to the ESSC. A cost analysis showed that this scenario can result in cost savings of more than \$200.000 per year in physical distribution cost.

Relation with the SPSS

This last case study addresses the issue of using different distribution configurations for different service part categories. In terms of flexibility defined in the SPSS this is again a form of allocation flexibility. The first part of the project investigated the possibility of direct shipments directly from the supplier to the warehouses in the European countries, eliminating the central European warehouse as a stocking location for certain parts. The second part of the project investigated the probability of centralizing service part inventories in the Netherlands toward the central European warehouse, eliminating the local warehouse in Gouda as stocking location for certain parts. All stocking locations in the SPSS are still necessary but, depending on service part and transport characteristics, some locations are skipped as stocking locations in the distribution configuration.

3.5.6 General findings

In retrospect we can identify two important issues that emerge from these case studies:

- Strategic importance of allocation flexibility.
- Choice of performance indicators.

In all cases under consideration we saw that the allocation of stock in the supply system is of strategic importance. For example, the Intergraph-case illustrates that significant cost savings can be obtained when centralizing parts inventories that need overnight supply anyway. The Digital-case

illustrates the potential benefits of vendor managed inventory: ship parts directly from the vendor to the national warehouses in Europe. In the Philip Morris case a decision support tool for the 0-1 decision was developed, taking into consideration several aspect such as criticality, availability, and cost. The strategic issue of determining the parts assortment at different stocking locations forms the basic question for service parts logistics in general.

A second observation that can be made from these case studies is the importance of the performance indicator that is being used. In several cases the traditional fill rate measure was used. However, other measures were used as well such as work order completion rate (Intergraph-case, Digital-case) and operational availability of the technical system that has to be maintained (ASML-case). Using different performance indicators can lead to different stocking policies for service parts. The choice of performance indicators is therefore very important in service part logistics.

3.6 Conclusions

The SPSS framework presented in this chapter can be used to identify opportunities to increase the flexibility in the distribution and repair network for service parts. When considering the implementation of one of these flexibility options in an SPSS in practice, a thorough (quantitative) trade-off analysis has to be made. In this way the pro's and con's (in terms of increasing service performance and increasing cost) can be balanced and a well-based decision can be taken. In the next three chapters we present two quantitative trade-off models and one simulation model that can be used to support the decision on using flexibility. In Chapter 4 we present the Emergency Repair Model that can be used to analyze the benefits of applying fast repair. In Chapter 5 we present the Emergency Supply Model that can be used in the same way to analyze the effects of applying flexible supply in a two-echelon distribution network. Finally, in Chapter 6 we present a simulation model for evaluating different policies for applying flexibility in a three-echelon distribution network.

Chapter 4

The Emergency Repair Model

4.1 Introduction

In the SPSS framework we identified several flexibility options that can be applied to increase the performance of the repair and supply processes of service parts in a distribution network. In this chapter we focus our attention to the repair processes of failed service parts at repair centers. The performance of the repair process of failed parts can be influenced by applying repair flexibility as identified in the framework (see figure 4.1). *Work order release flexibility* (e.g. release repair work orders to the shopfloor for those parts that are needed most, outsource repair work at an external repair shop when the workload at the own repair shop is too high) and *repair shop flexibility* (e.g. use priority scheduling in the repair shop, extend the repair shop capacity temporarily by means of overtime, apply flexible manpower planning as described in De Haas (1995)) are means to control the efficiency and effectiveness of the repair activities at repair centers. A trade-off analysis has to be made between the revenues and the costs of using these flexibilities.

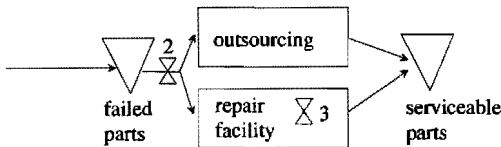


Figure 4.1: *Work order release flexibility (2) and repair shop flexibility (3) in the SPSS*

Two important research questions arise when discussing the expedition of the repair process of service parts:

- 1) *What policy should be used to trigger emergency repair?*
- 2) *What cost structure is needed for a trade-off analysis?*

The first research question addresses the conditions under which flexibility should be applied in order to speed up the repair process. Should the repair process for a specific service part be expedited when the considered inventory location is out of stock or when the physical stock on hand drops below a critical threshold value? The second research question addresses the cost structure that must be used for evaluating the efficiency and effectiveness of using this kind of flexibility. The

extra cost of using emergency procedures has to be balanced against the reduction of the inventory holding cost and the reduction of the penalty cost (as a result of reduced waiting time for backordered customers). Given a certain target service performance one has to make a trade off between inventory holding cost and using the emergency replenishment procedure. The key issue here is that we only consider those costs that are affected by the decision to implement repair flexibility.

In this chapter we present an analytic single-item single-echelon model with one-for-one replenishment, the Emergency Repair Model (ERM), that addresses these two questions in a repair shop environment (see Verrijdt *et al.*, 1996). Defective parts arriving at a repair shop can be sent into emergency repair instead of normal repair. In practice this could mean that the repair work is outsourced to an external repair facility. It could also mean that extra repair capacity is hired to expedite the repair work. De Haas (1995) for example investigates the use of flexible manpower planning in such a situation. In case of a multi-product situation (which we do not consider in our model), using priority rules is also an option to influence the repair throughput times. In our model parts are sent into an emergency repair channel when the net inventory of serviceable parts (i.e. physical stock on hand minus backorders) is equal to or lower than an emergency trigger level.

This chapter is organized as follows. In Section 4.2 we discuss some of the literature that considers the use of emergency repair (or emergency supply) for service parts. We focus our attention on some models that are closely related to our ERM and discuss the differences. In Section 4.3 the Emergency Repair Model is presented and the relevant service measures and cost structure are introduced. An analysis of the ERM is presented in Section 4.4. In contrast to most literature in this field that yield models with approximations of the operating characteristics, we present an *exact* analysis of the model. In Section 4.5 we present some numerical results with respect to service and cost performance of the ERM. In this section we also analyze the sensitivity of the model performance with respect to the choice of repair time distribution. In our model analysis we assume exponentially distributed repair times which allows us to use a Markovian analysis and to derive exact expressions for the cost and service performance. We prove that in two extreme cases the numerical results are only dependent on the mean of the repair distribution and not on the repair distribution itself (Section 4.4.4). One case is characterized by an emergency repair rate equal to the normal repair rate (i.e. no emergency repair) and the other case is characterized by an infinite emergency repair rate (i.e. instantaneous emergency repair). We use simulation to show a similar insensitivity result for any intermediate value of the emergency repair rate as well (Section 4.5.1). Furthermore, we compare our emergency repair policy with cost-optimal stock levels and emergency trigger levels with two other policies in terms of cost performance. These two policies are: 1) no emergency repair is used (i.e. one repair speed in all situations), and 2) the emergency trigger level is equal to zero (i.e. emergency repair is used when the physical stock on hand drops to zero). In

Section 4.6 we compare our cost results with the results of a model by Muckstadt and Thomas (1980). This latter model is discussed in more detail in the next section (see also Chapter 2). Finally, in Section 4.7 we discuss the model and its performance and draw some final conclusions.

4.2 Related literature

A number of papers consider the possibility of using alternative supply modes for stock replenishment. Some early models in the sixties describe periodic review base stock policies with two alternative lead times: one-period lead time for normal supply and zero-period lead time for emergency supply (see e.g. Barankin (1961), Daniel (1962), Fukuda (1964)). Whittmore and Saunders (1977) extend this analysis to situations with lead times of arbitrary length. Rosenshine and Obee (1976) compare the performance of a standing-order policy (i.e. a policy with fixed-size order arrivals at the beginning of each review period) allowing for emergency orders in case of shortages and allowing for sell-offs in case of stock surplus, with a traditional periodic review base stock policy. Moinzadeh and Nahmias (1988) analyze a continuous review (s,S) -inventory policy with different reorder points for normal and emergency replenishments. Cohen *et al.* (1988) also consider an (s,S) -inventory policy with two priority classes of customers: normal replenishment orders and emergency replenishment orders with higher priority. Ernst and Cohen (1993) extend this model to a situation where the classification of customers is a decision variable.

The inventory models used for controlling service part inventories are mostly continuous review $(S-1,S)$ -policies with Poisson demand. Such models are appropriate for low demand rates and expensive items such that ordering cost is negligible compared to holding cost. The METRIC-model by Sherbrooke (1968) is a well-known and often cited multi-echelon model in this category (see chapter 2 for a detailed description and discussion of the METRIC model). The METRIC model, however, does not allow for emergency shipments in stock-out situations. Muckstadt and Thomas (1980) extend the METRIC model to allow for such emergency shipments in a two-echelon inventory system (in a later paper Hausman and Erkip (1994) use this model for further analysis). Fast direct deliveries from the central warehouse or from the plant are used when local warehouses run out of stock, in order to minimize customer waiting time. The main purpose of their work is to show the cost benefit of using multi-echelon modelling techniques in comparison with single-echelon models for each separate inventory location. They do not explicitly investigate the trade off between emergency shipments and inventory investment. The reason why we compare their model (adapted for a single-echelon situation) to our ERM is that they assume that a backordered demand is always satisfied by the emergency order it invoked, even if a normal replenishment order arrives earlier. In our ERM we assume that a backordered demand is filled with the part that first becomes available from the repair process, either normal or emergency repair.

Moinzadeh and Schmidt (1991) present an approximate single-echelon model with two modes of resupply. If the physical stock on hand drops to a certain threshold value and the remaining lead time for a pipeline order exceeds the lead time for an emergency order, then an emergency replenishment order is issued. In order to use information about pipeline orders, they assume constant replenishment times. They model both the backordering and the lost demand situation and present a technique for calculating the optimal stock level and trigger level that minimize a cost function. Because of the exponential repair time assumption in the ERM we can not, in contrast to Moinzadeh and Schmidt, check the pipeline for normal repair orders that would arrive before a newly issued emergency order. The fact that we cannot check the pipeline for early normal replenishment order arrivals in case of a stock-out, implies that in our model the number of emergency orders is higher. On the other hand, if a backordered demand is filled by an early normal replenishment order arrival, the emergency order that was issued will increase the service performance for *future* customers. Consequently, our ERM will show higher emergency replenishment cost but lower penalty cost. Therefore the cost structure of the model (emergency repair costs versus penalty costs) determines the (dis-) advantage of checking the pipeline.

Finally, we discuss a paper by Aggarwal and Moinzadeh (1994) who present an approximate two-echelon model where retail centers are supplied by a central production facility that produces to order. The retail centers can issue normal and emergency resupply orders and the emergency orders have priority over the normal orders at the production facility. The retailers apply an $(S-I, S)$ inventory policy and the production facility is modelled as an $M/G/1$ queuing system. The authors show how to derive 'optimal' stock levels and trigger levels for emergency orders at the retail centers. Similar to our model, they do not use information about the pipeline orders when issuing emergency replenishment orders. Different policies are evaluated and the results are in line with the results of Moinzadeh and Schmidt (1991) for a single-echelon situation. In contrast to the model by Aggarwal and Moinzadeh, we model the emergency repair process as being independent of the normal repair process. This represents the situation in which emergency repair is outsourced to external repair centers and therefore does not interfere with the normal repair work.

In practice emergency repair is often used when stock on hand of serviceable parts is depleted, i.e. the emergency trigger level is set equal to zero. However, earlier research (e.g. Moinzadeh and Schmidt (1991) and Aggarwal and Moinzadeh (1994)) has shown that it can be very beneficial to use emergency repair when the net inventory is positive. In this way it is possible to anticipate on possible stock-out situations in the near future. In our ERM we also allow for arbitrary trigger levels that initiate emergency repair. It is even possible to allow for negative emergency trigger levels.

4.3 Model description

We consider a stocking location for service parts with initial stock level S where failed parts arrive according to a Poisson process with rate λ . An $(S-1, S)$ repair policy is applied. A failed part that arrives at the stocking location is exchanged for a serviceable part if available, otherwise a backorder is created. If the net inventory of serviceable parts after exchange exceeds a given emergency trigger level x , the failed part is sent to the repair shop for normal repair. If the net inventory of serviceable parts after exchange is equal to or lower than x , the part is sent into emergency repair. This alternative emergency repair process is independent of the normal repair process. A backorder is filled with the first service part that becomes available (either from the normal repair process or the emergency repair process) according to the FCFS-principle (First Come First Serve). The repair times (both normal and emergency) are assumed to be exponentially distributed. We assume that a failed part can be immediately taken into normal repair (if the net inventory exceeds level x) or emergency repair (if the net inventory is equal to or lower than level x) and that no waiting time occurs. The same modelling assumption of the repair process is used in the well-known METRIC model (Sherbrooke, 1968). This infinite repair capacity assumption enables Sherbrooke to apply Palm's theorem. As a consequence, the operating characteristics of the METRIC model only depend on the mean repair time and are independent of the repair time distribution itself. In Section 4.5 we address this insensitivity issue for our ERM. The model is depicted in figure 4.1. Note that it is possible to allow for negative trigger levels in our model. For example, a trigger level $x = -1$ implies that emergency repair orders are only issued when at least one backorder exists.

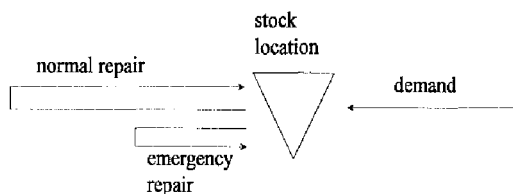


Figure 4.2: The Emergency Repair Model

We use the following notation:

- λ : Poisson arrival rate of failed parts at the stocking facility
- μ : repair rate for parts sent into normal repair (exponentially distributed)
- τ : repair rate for parts sent into emergency repair ($\tau \geq \mu$, exponentially distributed)
- S : initial stock level of service parts

- x : emergency trigger level
 (i,j) : system state with i parts in normal repair ($0 \leq i \leq S-x$) and j parts in emergency repair ($j \geq 0$)
 p_{ij} : steady state probability associated with state (i,j)

In Section 4.4 we show how to calculate the steady state probabilities p_{ij} using Markov chain analysis. These probabilities enable us to compute a number of performance measures for our model. We define the following operating characteristics:

- L^e : expected number of parts in emergency repair
 L^n : expected number of parts in normal repair
 L : expected number of parts in repair
 B : expected number of backorders

Then:

$$\begin{aligned}
 L^e &= \sum_{i=0}^{S-x} \sum_{j=1}^{\infty} j p_{ij} \\
 L^n &= \sum_{i=1}^{S-x} \sum_{j=0}^{\infty} i p_{ij} \\
 L &= L^e + L^n = \sum_{k=1}^{\infty} k r(k) \quad , \quad r(k) = \sum_{i+j=k} p_{ij} \\
 B &= \sum_{k=S+1}^{\infty} (k-S) r(k) = L - S + \sum_{k=0}^S (S-k) r(k)
 \end{aligned}$$

Service performance measures

We derive expressions for three service measures:

- 1) FR : Fill rate
- 2) W^* : Average response time (or waiting time) for a backorder
- 3) $v(t)$: Probability that the response time for a backorder exceeds a time limit t

The fill rate is defined as the fraction of demand that can be satisfied from stock on hand. By the PASTA property (Poisson Arrivals See Time Average, see Wolff, 1982) this is the same as the fraction of time there are at most $S-1$ parts in repair:

$$FR = \sum_{i+j < S} p_{ij} \quad (4.1)$$

Emergency Repair Model

The average response time W^* for a backorder is an important service measure in case of high penalty cost. Let W represent the expected waiting time for an arbitrary part to be exchanged by a serviceable part:

$$\begin{aligned} W &= Pr\{stock\ on\ hand = 0\} W^* + Pr\{stock\ on\ hand > 0\} 0 \\ &= (1 - FR) W^* \end{aligned}$$

Using Little's formula (i.e. $W = B / \lambda$) we now have the following expression for the expected duration (or response time) W^* of a backorder:

$$W^* = \frac{B}{\lambda (1 - FR)} \quad (4.2)$$

The probability that the response time W for a backorder exceeds a contractually agreed time limit is an important indicator for measuring the variation in the response time. By the PASTA property this probability can be calculated as follows:

$$v(t) = Pr(W > t) = \sum_{i \geq 25} Pr(W_{i,j} > t) p_{i,j} \quad (4.3)$$

where $W_{i,j}$ represents the waiting time for a part that arrives when the system is in state (i,j) . A detailed elaboration of this expression can be found in Appendix B.

Cost structure

An important aspect to be considered when discussing the trade off between inventory investment and emergency repair is the cost structure of the model. We consider the following four operational cost factors (see also Chapter 1): inventory holding cost (C_1), emergency repair cost (C_2), normal repair cost (C_3), and penalty cost per time unit for backorders (C_4). The total cost function TC can be expressed as follows:

$$TC = C_1 + C_2 + C_3 + C_4 \quad (4.4)$$

with:

$$C_1 = h E[\text{inventory on hand}] = h \sum_{k=1}^S k \sum_{i+j=S-k} p_{ij}$$

$$C_2 = Pr(\text{emergency repair}) \lambda f^e K = \sum_{i+j \geq S-x} p_{ij} \lambda f^e K$$

$$C_3 = Pr(\text{normal repair}) \lambda f^n K = \sum_{i+j < S-x} p_{ij} \lambda f^n K$$

$$C_4 = Pr(\text{stock out}) \lambda W^* p = (1-FR) \lambda W^* p = B p$$

where

- K : unit price of one service part
 h : annual holding cost for one service part (as fraction of K)
 f^e : emergency repair cost for one service part (as fraction of K)
 f^n : normal repair cost for one service part (as fraction of K)
 p : penalty cost per time unit per part backordered

Note that when the trigger level is equal to zero, the probability that a failed part is sent into normal repair is equal to the fill rate. The cost of emergency repair (f^e) typically depends on the speed of the emergency repair (τ). For ease of presentation we assume a linear relationship between f^e and τ (see figure 4.3). In practice the relation between cost and speed of emergency repair is often more complex. For example, when emergency repair work is outsourced to an external vendor, the emergency repair cost reflects the effort that has to be made by the vendor. The price that has to be paid is usually the result of negotiations between outsourcer and vendor and may vary per situation. Here we assume a linear relation. For $\tau = \mu$ (i.e. emergency repair is as fast as normal

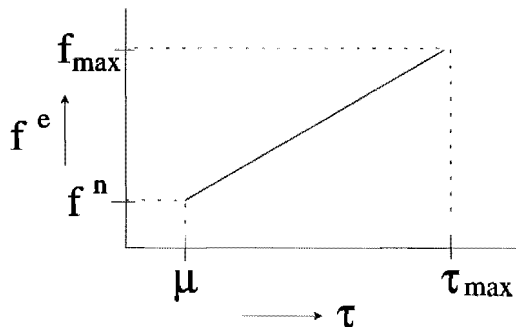


Figure 4.3: Relation between speed and cost of emergency repair

Emergency Repair Model

repair) the emergency repair cost is equal to the normal repair cost ($f^e = f^n$). For $\tau = \tau_{\max}$ the emergency repair cost is equal to a maximum value ($f^e = f_{\max}$). τ_{\max} represents a maximum emergency repair rate. Note that the value of f_{\max} is an indication of the cost of using emergency repair. The emergency repair cost for any value of τ ($\mu \leq \tau \leq \tau_{\max}$) can be calculated as follows:

$$f^e = \frac{f_{\max} - f^n}{\tau_{\max} - \mu} \tau + f^n - \mu \frac{f_{\max} - f^n}{\tau_{\max} - \mu} \quad (4.5)$$

4.4 Model analysis

In this section we show how the steady state probabilities $p_{i,j}$ can be calculated. Approximations for $p_{i,j}$ can be obtained by truncating the state space and then numerically solving the remaining finite Markov chain. Disadvantages of this approach are that in many cases it is numerically expensive and no guarantee can be given for the accuracy of the solution. Here we will present an *exact* analysis. First we consider the Markov process embedded on the states (i,j) with $0 \leq i \leq S-x$, $0 \leq j \leq S-x$ and compute the associated steady state probabilities $\tilde{p}_{i,j}$ (Section 4.4.1). The problem here is to identify the transition rates of the embedded process. Usually this is as difficult as finding the probabilities $p_{i,j}$ of the original problem. However, it appears that for the present problem, we are able to determine explicit expressions for these rates (Section 4.4.2). Next, we return to the original Markov process and compute the steady state probabilities $p_{i,j}$ (Section 4.4.3). We present two limiting cases in which the performance of the ERM is independent of the choice of repair time distribution, but only depends on the mean repair time (Section 4.4.4). Finally, we show how the present model can be extended to model the situation with *one* normal repair channel (Section 4.4.5).

4.4.1 The embedded Markov process

The transition rate diagram of the ERM is depicted in figure 4.4. We define level j as the set of all states with j parts in emergency repair:

$$\text{level } j := \{ (i,j) \mid 0 \leq i \leq S-x \}$$

A natural approach to analyze the ERM is to partition the state space into levels 0,1,2... and then try to apply Neuts' matrix geometric approach (Neuts, 1981). This approach requires that the transition rates are constant from some level onwards. In the present model, however, the rates are

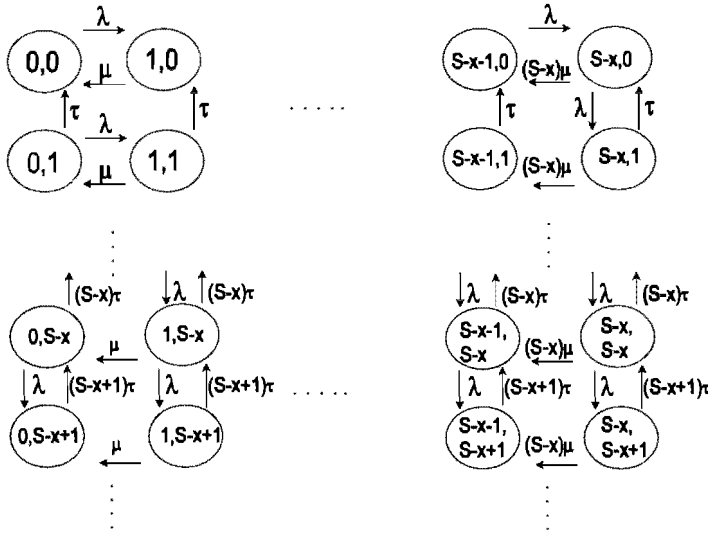


Figure 4.4: Transition rate diagram of the ERM

level dependent. Therefore, the matrix geometric approach cannot be used, although recently some progress has been made in extending this approach to models with level dependent transition rates (see Ramaswami and Taylor, 1996). Instead we proceed as follows. We first consider the process embedded on the states (i, j) with $0 \leq i \leq S-x, 0 \leq j \leq S-x$ (i.e. excursions to levels higher than $S-x$ are not considered). Note that when the system is in a state at level $S-x$, an arriving failed part will always be sent into emergency repair. The problem is to calculate the transition rates for the states $(i, S-x)$ in the embedded Markov process, corresponding to an excursion of the original process. In state $(i, S-x)$ an excursion to higher levels starts with rate λ and eventually the process will return to level $S-x$ in some state $(k, S-x)$ with $k \leq i$. Define $\pi_{i,k}$ as the probability that, given the original process starts in state $(i, S-x+1)$ at level $S-x+1$, it returns for the first time to level $S-x$ in state $(k, S-x)$. Note that $\pi_{i,k} = 0$ for $k > i$. Then the transition rate from state $(i, S-x)$ to state $(k, S-x)$ corresponding to an excursion is equal to $\lambda \pi_{i,k}$. The transition rate diagram of the embedded Markov process is depicted in figure 4.5.

Let $\bar{p}_{i,j}$ represent the steady state probabilities of the embedded Markov process. The equilibrium probabilities $\bar{p}_{i,j}$ for the embedded system can now be computed by solving the (finitely many) equilibrium equations, together with the normalization constraint:

$$\sum_{i=0}^{S-x} \sum_{j=0}^{S-x} \tilde{p}_{i,j} = 1$$

But to do so, we must first calculate the transition probabilities $\pi_{i,k}$.

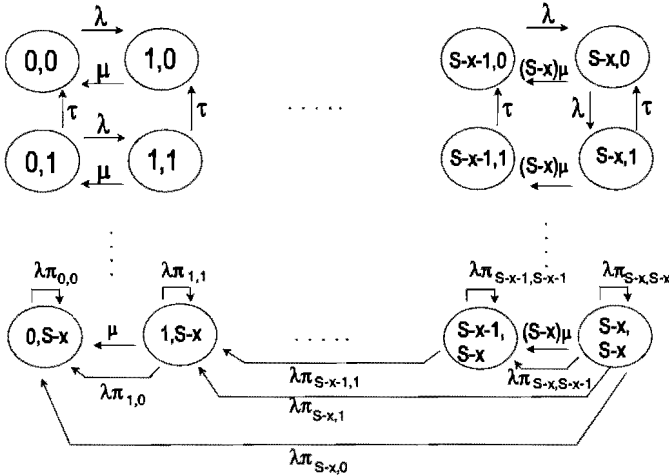


Figure 4.5: Transition rate diagram of the embedded model

4.4.2 Calculation of the transition probabilities $\pi_{i,k}$

During an excursion of the original Markov process to the levels higher than $S-x$, arriving failed parts are all sent into the emergency repair channel, and hence the system acts as an $M/M/\infty$ queue. Let us define T_n as the time needed in an $M/M/\infty$ queue with arrival rate λ and service rate τ , to bring down the number of customers from n to $n-1$ (Busy Period Analysis). It then follows that the time the original Markov process needs to return to level $S-x$ when starting in level $S-x+1$, is stochastically identical to T_{S-x+1} .

Suppose that at time $t=0$ the system is in level $S-x+1$ and i parts are in normal repair. Then $\pi_{i,k}$ represents the probability that k of these parts ($k \leq i$) are repaired in time T_{S-x+1} , which is the length of the excursion. Define:

$$r_{k,i}(t) := \Pr\{k \text{ parts are repaired in the normal repair channel in time } t \mid i \text{ parts are in normal repair at time } t=0\}$$

By conditioning on T_{S-x+1} we find:

$$\pi_{i,i-k} = \int_0^{\infty} r_{k,i}(t) dPr\{T_{S-x+1}=t\} \quad (4.6)$$

Using the fact that the normal repair times are exponentially distributed with parameter μ and application of the binomium of Newton gives:

$$\begin{aligned} r_{k,i}(t) &= \binom{i}{k} (1 - e^{-\mu t})^k (e^{-\mu t})^{i-k} \\ &= \binom{i}{k} \sum_{n=0}^k \binom{k}{n} (-1)^{k-n} e^{-\mu(i-n)t} \end{aligned}$$

Substitution of this relation in (4.6) gives:

$$\begin{aligned} \pi_{i,i-k} &= \binom{i}{k} (-1)^k \sum_{n=0}^k \binom{k}{n} (-1)^n \int_0^{\infty} e^{-\mu(i-n)t} dPr\{T_{S-x+1}=t\} \\ &= \binom{i}{k} (-1)^k \sum_{n=0}^k \binom{k}{n} (-1)^n \varphi_{S-x+1}(\mu(i-n)) \end{aligned}$$

where $\varphi_n(s)$ represents the Laplace-Stieltjes transform of T_n . The Laplace-Stieltjes transform $\varphi_n(s)$ can be calculated from the following expressions (see Appendix C for a detailed analysis):

$$\begin{aligned} \varphi_1(s) &= \frac{1}{\lambda} \left\{ \lambda + s - \frac{s}{1 + \sum_{k=1}^{\infty} \prod_{i=1}^k \frac{-\lambda}{s+i\tau}} \right\} , \\ \varphi_{n+1}(s) &= \frac{(\lambda + n\tau + s) \varphi_n(s) - n\tau}{\lambda \varphi_n(s)} , \quad n=1,2,\dots \end{aligned}$$

4.4.3 Calculation of $p_{i,j}$

For states (i,j) with $j \leq S-x$ the probabilities $p_{i,j}$ of the original process follow from

Emergency Repair Model

$$p_{ij} = \tilde{p}_{ij} \cdot C \quad , \quad 0 \leq i \leq S-x, \quad 0 \leq j \leq S-x \quad ,$$

where C is the (unknown) probability that the original process is in the set of states (i, j) with $0 \leq i \leq S-x$, $0 \leq j \leq S-x$:

$$C = \sum_{i=0}^{S-x} \sum_{j=0}^{S-x} p_{ij}$$

The probability that the (embedded) system is in a state at level j is defined as follows:

$$q(j) := \sum_{i=0}^{S-x} p_{ij} \quad , \quad j \geq 0$$

$$\tilde{q}(j) := \sum_{i=0}^{S-x} \tilde{p}_{ij} \quad , \quad 0 \leq j \leq S-x$$

$q(j)$ and $\tilde{q}(j)$ represent the probability that j parts are in emergency repair (level j) in the original process and the embedded process, respectively. From the transition-rate diagram of the original process (figure 4.4) we can derive the following equation by balancing the flow between the levels $j-1$ and j with $j > S-x$:

$$\lambda q(j-1) = j \tau q(j)$$

so

$$q(j) = \frac{\lambda}{\tau} \frac{1}{j} q(j-1)$$

Repeated application of this relation yields for $j > S-x$:

$$\begin{aligned} q(j) &= \left(\frac{\lambda}{\tau} \right)^{j-S-x} \frac{1}{j(j-1) \dots (S-x+1)} q(S-x) \\ &= \left(\frac{\lambda}{\tau} \right)^j \frac{1}{j!} (S-x)! \left(\frac{\tau}{\lambda} \right)^{S-x} q(S-x) \\ &= \left(\frac{\lambda}{\tau} \right)^j \frac{1}{j!} (S-x)! \left(\frac{\tau}{\lambda} \right)^{S-x} C \tilde{q}(S-x) \end{aligned}$$

The constant C can now be computed as follows:

$$\begin{aligned}
 1 &= \sum_{j=0}^{\infty} q(j) \\
 &= \sum_{j=0}^{S-x} q(j) + \sum_{j=S-x+1}^{\infty} q(j) \\
 &= C \sum_{j=0}^{S-x} \bar{q}(j) + \sum_{j=S-x+1}^{\infty} \left(\frac{\lambda}{\tau}\right)^j \frac{1}{j!} (S-x)! \left(\frac{\tau}{\lambda}\right)^{S-x} C \bar{q}(S-x) \\
 &= C \left\{ 1 + (S-x)! \left(\frac{\tau}{\lambda}\right)^{S-x} \bar{q}(S-x) \left[e^{\lambda\tau} - \sum_{j=0}^{S-x} \left(\frac{\lambda}{\tau}\right)^j \frac{1}{j!} \right] \right\}
 \end{aligned}$$

so

$$C = \left\{ 1 + (S-x)! \left(\frac{\tau}{\lambda}\right)^{S-x} \bar{q}(S-x) \left[e^{\lambda\tau} - \sum_{j=0}^{S-x} \left(\frac{\lambda}{\tau}\right)^j \frac{1}{j!} \right] \right\}^{-1}$$

The steady state probabilities p_{ij} for $j > S-x$ can now be computed as follows. The equilibrium equation for state $(i, S-x)$ in the original Markov process can be used to calculate $p_{i, S-x+1}$:

$$\begin{aligned}
 p_{0, S-x+1} &= \frac{p_{0, S-x}((S-x)\tau + \lambda) - \mu p_{1, S-x}}{(S-x+1)\tau} , \\
 p_{i, S-x+1} &= \frac{p_{i, S-x}((S-x)\tau + \lambda + i\mu) - (i+1)\mu p_{i+1, S-x} - \lambda p_{i, S-x-1}}{(S-x+1)\tau} , & 0 < i < S-x , \\
 p_{S-x, S-x+1} &= \frac{p_{S-x, S-x}((S-x)\tau + \lambda + (S-x)\mu) - \lambda p_{S-x, S-x-1}}{(S-x+1)\tau} .
 \end{aligned}$$

Finally we can calculate p_{ij} recursively for all $j > S-x$:

$$\begin{aligned}
 p_{i, j+1} &= \frac{p_{i, j}(j\tau + \lambda + i\mu) - (i+1)\mu p_{i+1, j} - \lambda p_{i, j-1}}{(j+1)\tau} , & 0 \leq i < S-x , \\
 p_{S-x, j+1} &= \frac{p_{S-x, j}(j\tau + \lambda + (S-x)\mu) - \lambda p_{S-x, j-1}}{(j+1)\tau} .
 \end{aligned}$$

4.4.4 Limiting cases

The analysis of the ERM is based on the assumption of exponentially distributed repair times for both normal and emergency repair. Here we show that in two limiting cases of the ERM the choice of repair time distribution does not affect the service performance. Let ω_k denote the steady state probability of k parts in repair (either normal or emergency repair). Then there are two extreme situations in which the steady state probabilities ω_k (and therefore also the performance measures) are independent of the repair time distribution and only depend on the mean repair times.

Situation 1: When the emergency repair rate is equal to the normal repair rate (i.e. $\tau=\mu$), the ERM reduces to an $M(\lambda)M(\mu)\infty$ system with service rate μ . For such a queuing system we have the well known result by Palm (1938) stating that the number of customers in the system (i.e. parts in repair) is Poisson distributed and only depends on the mean of the repair time distribution:

$$\omega_k = e^{-\lambda/\mu} \frac{(\lambda/\mu)^k}{k!}$$

Situation 2: When the emergency repair rate τ goes to infinity (i.e. emergency repair is instantaneous), the ERM reduces to an $M(\lambda)M(\mu)S-x|S-x$ loss system. Failed parts that arrive when stock on hand is equal to or less than $S-x$ are repaired in time zero. Again, the steady state probabilities ω_k ($k=0..S-x$) only depend on the mean of the repair time distribution:

$$\omega_k = \frac{(\lambda/\mu)^k/k!}{\sum_{i=0}^s (\lambda/\mu)^i/i!}$$

We can conclude that in the two extreme situations described above ($\tau=\mu$ and $\tau=\infty$) the repair time distribution only affects the service performance through its mean value. In the next section we will test, by means of simulation, if this strong insensitivity result holds for intermediate values of τ as well.

4.4.5 Limited repair capacity

In the ERM the repair times of failed parts are independently distributed stochastic variables. This is because the normal repair process is modelled as an $MM|S-x$ queue and the emergency repair process is modelled as an $MM|\infty$ queue. Since the number of parts in normal repair is equal to or

less than $S-x$, the repair times of parts in normal repair are independent of each other. The assumption of independent repair times for parts in emergency repair is quite realistic, because the emergency repair process is often represented by outsourcing repair work to an external repair facility. In that situation one often makes contractual agreements with respect to the repair lead time of outsourced repair work. The normal repair process on the other hand, often represented by an internal repair shop, usually has limited repair capacity. In the ERM we also assume that the normal repair capacity is limited, i.e. the number of normal repair channels is equal to $S-x$. When all $S-x$ normal repair channels are occupied, arriving failed parts are sent into emergency repair and hence the normal repair times are still independent of each other. However, the Emergency Repair Model we presented before can be extended in such a way that the normal repair process is represented by an $M/M/1$ queue. In that case (assuming that $S-x$ is greater than one) normal repair lead times will be correlated. In the remainder of this section we show how the preceding analysis can be extended to the situation with one normal repair channel.

The transition rate diagrams of the ERM and the embedded model for the situation with one normal repair channel are almost identical to the ones presented in Section 4.4.1. The only difference is that all the normal repair transition rates $j\mu$ ($j=1..S-x$) have to be replaced by μ , since only one normal repair channel is available. Furthermore, the calculation of the transition probabilities $\pi_{i,k}$ changes as well. In Section 4.4.2 we defined $r_{k,i}(t)$ as the probability that k parts are repaired in the normal repair channel in time t , assuming that there were i parts in normal repair at time zero. In the situation with one normal repair channel, these probabilities are now calculated as follows:

$$r_{k,i}(t) = e^{-\mu t} \frac{(\mu t)^k}{k!}, \quad 0 \leq k < i$$

$$r_{i,i}(t) = 1 - \sum_{k=0}^{i-1} r_{k,i}(t)$$

The probability that k ($<i$) parts are repaired in time t is equal to the Poisson probability of exactly k arrivals in time t with arrival rate μ . The probability that all i parts are repaired in time t is equal to the Poisson probability of at least i arrivals in time t with arrival rate μ .

Using expression (4.6), we can now calculate the transition probabilities $\pi_{i,i-k}$ ($k<i$) as follows:

$$\begin{aligned}
 \pi_{i,j-k} &= \int_0^{\infty} r_{kj}(t) dPr(T_{S-x+1}=t) \\
 &= \int_0^{\infty} e^{-\mu t} \frac{(\mu t)^k}{k!} dPr(T_{S-x+1}=t) \\
 &= \frac{(-\mu)^k}{k!} \int_0^{\infty} (-t)^k e^{-\mu t} dPr(T_{S-x+1}=t) \\
 &= \frac{(-\mu)^k}{k!} \frac{d^k}{ds^k} \{ \varphi_{S-x+1}(s) \}_{s=\mu}
 \end{aligned}$$

The recursive relation between the Laplace-Stieltjes transforms $\varphi_{n+1}(s)$ and $\varphi_n(s)$, derived in Appendix B, can be used to calculate the k -th derivative of $\varphi_{S-x+1}(s)$, and hence the transition probabilities $\pi_{i,j-k}$ can be determined (see Appendix D for a detailed analysis).

Given the adjusted transition probabilities and the adjusted normal repair rates in the diagram, we can calculate the performance characteristics of the model in a similar way as described before.

4.5 Numerical evaluation

In this section we present some numerical results for the service and cost performance of our model. In Section 4.5.1 we analyze the service performance with respect to the fill rate and the average response time for a backorder. We restrict ourselves to the situation where the trigger level is equal to zero, in order to investigate the impact of emergency repair on these service performance measures. We also analyze the sensitivity of the service performance with respect to the choice of repair time distribution. We compare the service performance of the ERM with the service performance of a simulation model in which *deterministic* repair lead times are assumed. In Section 4.5.2 we address the issue of multi-criteria analysis. The selection of an initial stock level in combination with an emergency repair rate for a given situation, depends on the strategic goal that is defined. In Section 4.5.3 we analyze the cost behavior of our model and we calculate the initial stock level S and the trigger level x that minimize the total cost for a given emergency repair rate. We compare these results with the results from the model in which no emergency repair is allowed (i.e. $\tau = \mu$) and the model in which emergency orders are only issued when the net inventory is zero (i.e. $x = 0$). Finally, in Section 4.5.4 we elaborate on the sensitivity analysis of the numerical results with respect to the model assumptions. More specifically, we use simulation to test the sensitivity of the fill rate performance of the model with respect to the assumption of Poisson demand arrivals and exponentially distributed repair lead times.

4.5.1 Service performance and sensitivity analysis

Here we analyze the service performance of the Emergency Repair Model. We focus on the fill rate (FR) and the average response time for a backorder (W^*). These two service measures are calculated and presented in table 4.1 and table 4.2. Note that W^* is normalized by multiplying it with the normal repair rate μ . $W^*\mu$ represents the expected backorder duration as a fraction of the mean

| ρ | S | τ/μ | ERM | | SIMULATION | |
|--------|---|------------|-------|----------|------------|----------|
| | | | FR | $W^*\mu$ | FR | $W^*\mu$ |
| 0.1 | 0 | 1 | 0.000 | 1.000 | 0.000 | 1.000 |
| | | 5 | 0.000 | 0.200 | 0.000 | 0.200 |
| | | 10 | 0.000 | 0.100 | 0.000 | 0.100 |
| | 1 | 1 | 0.905 | 0.508 | 0.905 | 0.508 |
| | | 5 | 0.909 | 0.167 | 0.909 | 0.180 |
| | | 10 | 0.909 | 0.091 | 0.909 | 0.095 |
| | 2 | 1 | 0.995 | 0.339 | 0.995 | 0.340 |
| | | 5 | 0.995 | 0.143 | 0.995 | 0.164 |
| | | 10 | 0.995 | 0.083 | 0.995 | 0.091 |
| | 3 | 1 | 1.000 | 0.254 | 1.000 | 0.241 |
| | | 5 | 1.000 | 0.125 | 1.000 | 0.141 |
| | | 10 | 1.000 | 0.077 | 1.000 | 0.085 |
| 0.5 | 0 | 1 | 0.000 | 1.000 | 0.000 | 1.000 |
| | | 5 | 0.000 | 0.200 | 0.000 | 0.200 |
| | | 10 | 0.000 | 0.100 | 0.000 | 0.100 |
| | 1 | 1 | 0.607 | 0.541 | 0.607 | 0.542 |
| | | 5 | 0.663 | 0.166 | 0.665 | 0.180 |
| | | 10 | 0.666 | 0.091 | 0.667 | 0.095 |
| | 2 | 1 | 0.910 | 0.362 | 0.910 | 0.363 |
| | | 5 | 0.922 | 0.143 | 0.922 | 0.162 |
| | | 10 | 0.923 | 0.083 | 0.923 | 0.090 |
| | 3 | 1 | 0.986 | 0.270 | 0.986 | 0.269 |
| | | 5 | 0.987 | 0.125 | 0.987 | 0.145 |
| | | 10 | 0.987 | 0.077 | 0.987 | 0.086 |
| | 4 | 1 | 0.998 | 0.214 | 0.998 | 0.209 |
| | | 5 | 0.998 | 0.111 | 0.998 | 0.135 |
| | | 10 | 0.998 | 0.071 | 0.998 | 0.083 |
| | 5 | 1 | 1.000 | 0.177 | 1.000 | 0.166 |
| | | 5 | 1.000 | 0.100 | 1.000 | 0.130 |
| | | 10 | 1.000 | 0.067 | 1.000 | 0.066 |

Table 4.1: Fill rate (FR) and response time ($W^*\mu$) performance for $\rho=0.1$ and $\rho=0.5$

| ρ | S | τ/μ | ERM | | SIMULATION | |
|--------|----|------------|-------|----------|------------|----------|
| | | | FR | $W^*\mu$ | FR | $W^*\mu$ |
| 1 | 0 | 1 | 0.000 | 1.000 | 0.000 | 1.000 |
| | | 5 | 0.000 | 0.200 | 0.000 | 0.200 |
| | | 10 | 0.000 | 0.100 | 0.000 | 0.100 |
| | 1 | 1 | 0.368 | 0.582 | 0.369 | 0.582 |
| | | 5 | 0.491 | 0.166 | 0.495 | 0.179 |
| | | 10 | 0.498 | 0.091 | 0.499 | 0.095 |
| | 2 | 1 | 0.736 | 0.392 | 0.736 | 0.392 |
| | | 5 | 0.794 | 0.143 | 0.796 | 0.162 |
| | | 10 | 0.798 | 0.083 | 0.799 | 0.090 |
| | 3 | 1 | 0.920 | 0.291 | 0.920 | 0.290 |
| | | 5 | 0.935 | 0.125 | 0.935 | 0.147 |
| | | 10 | 0.937 | 0.077 | 0.937 | 0.086 |
| | 4 | 1 | 0.981 | 0.229 | 0.981 | 0.228 |
| | | 5 | 0.984 | 0.112 | 0.984 | 0.134 |
| | | 10 | 0.984 | 0.071 | 0.984 | 0.082 |
| | 5 | 1 | 0.996 | 0.188 | 0.996 | 0.187 |
| | | 5 | 0.997 | 0.101 | 0.997 | 0.122 |
| | | 10 | 0.997 | 0.067 | 0.997 | 0.078 |
| 6 | 1 | 0.999 | 0.159 | 0.999 | 0.156 | |
| | 5 | 0.999 | 0.092 | 0.999 | 0.119 | |
| | 10 | 0.999 | 0.063 | 0.999 | 0.076 | |
| 2 | 0 | 1 | 0.000 | 1.000 | 0.000 | 1.000 |
| | | 5 | 0.000 | 0.200 | 0.000 | 0.200 |
| | | 10 | 0.000 | 0.100 | 0.000 | 0.100 |
| | 1 | 1 | 0.135 | 0.657 | 0.136 | 0.657 |
| | | 5 | 0.318 | 0.165 | 0.324 | 0.178 |
| | | 10 | 0.329 | 0.091 | 0.331 | 0.095 |
| | 2 | 1 | 0.406 | 0.456 | 0.407 | 0.456 |
| | | 5 | 0.583 | 0.142 | 0.588 | 0.161 |
| | | 10 | 0.595 | 0.083 | 0.597 | 0.090 |
| | 3 | 1 | 0.677 | 0.337 | 0.677 | 0.337 |
| | | 5 | 0.777 | 0.125 | 0.779 | 0.146 |
| | | 10 | 0.785 | 0.077 | 0.787 | 0.085 |
| | 4 | 1 | 0.857 | 0.263 | 0.857 | 0.262 |
| | | 5 | 0.898 | 0.112 | 0.899 | 0.134 |
| | | 10 | 0.902 | 0.071 | 0.903 | 0.082 |
| | 5 | 1 | 0.947 | 0.214 | 0.947 | 0.213 |
| | | 5 | 0.960 | 0.101 | 0.960 | 0.123 |
| | | 10 | 0.962 | 0.067 | 0.962 | 0.078 |
| 6 | 1 | 0.983 | 0.179 | 0.983 | 0.180 | |
| | 5 | 0.987 | 0.093 | 0.987 | 0.114 | |
| | 10 | 0.987 | 0.062 | 0.988 | 0.075 | |

Table 4.2: Fill rate (FR) and response time ($W^*\mu$) performance for $\rho=1$ and $\rho=2$

normal repair time. Three parameters are varied: the utilization rate ρ ($=\lambda/\mu$); the initial stock level S ; and the relative speed of an emergency repair (τ/μ). The trigger level x is set to zero in all cases. Both the ERM-results (with exponentially distributed repair lead times) and the simulation results (with deterministic repair lead times) are presented. The analytic results (ERM) show that an increase in FR can be realized by using (faster) emergency repair. See for example the case with $\rho = 2$ and $S = 4$. The FR increases from 0.857 to 0.902 when an emergency repair channel is used that is ten times faster than normal repair. Note that the highest increase in FR is obtained for high values of ρ . With respect to the normalized duration of a backorder, similar observations can be made. When aiming for a target normalized backorder duration, there are different ways to realize that goal. For example when $\rho = 0.5$, the combinations $(S, \tau/\mu) = (0, 10)$ and $(S, \tau/\mu) = (5, 5)$ give the same normalized backorder duration of 10%. In that case, the cost structure of the system determines the optimal (i.e. cheapest) solution. In general we see that a significant reduction of the backorder duration can be obtained when using fast emergency repair. The main conclusion that can be drawn from the results in table 4.1 and table 4.2 is that a pre-specified service performance (in terms of fill rate or backorder response time) can be obtained in various ways. The trade-off between inventory cost, emergency repair cost, and penalty cost determine the solution that is cost-optimal. In Section 4.5.2 we address this cost trade-off and evaluate different policies. Note that in table 4.1 and table 4.2 the trigger level x is always equal to zero. Allowing non-zero trigger levels can also influence the performance of the model. Therefore, the trigger level can be considered a decision variable as well (see Section 4.5.2).

In Section 4.4.4 we proved that the service performance of the ERM is only dependent on the mean of the repair time distribution and independent of the repair time distribution itself in two extreme situations: 1) no emergency repair ($\tau = \mu$) and 2) instantaneous emergency repair ($\tau = \infty$). In these two situations the service performance of the model only depends on the mean normal repair time $1/\mu$. In order to test the sensitivity of the model performance, with respect to the choice of repair time distribution, for intermediate values for τ ($\mu < \tau < \infty$) as well, we simulated the ERM with deterministic repair times. A failed part that is sent into repair (either normal or emergency repair) is returned into serviceable state after a constant time interval. Especially for emergency repair, deterministic repair times are more realistic than exponentially distributed repair times. That is because emergency repair orders can be performed by external repair shops at fixed repair lead times. The simulation results with deterministic repair times are tabulated in tables 4.1 and 4.2. In each case 500.000 part arrivals were simulated. The relative deviations between the simulation results and the analytic results are summarized in table 4.3. The minimum, the maximum, and the average of the relative deviations over all 72 cases (from tables 4.1 and 4.2) are shown. Especially the fill rate performance is very close to the analytic (ERM) results. This important insensitivity result indicates that the assumption of exponentially distributed repair times is not very restrictive for our ERM.

| SIMULATION PERFORMANCE | SERVICE MEASURE: | |
|------------------------|------------------|----------|
| | FR | $W^*\mu$ |
| minimum deviation | 0.0 % | 0.0 % |
| maximum deviation | 1.9 % | 30.0 % |
| average deviation | 0.1 % | 8.0 % |

Table 4.3: Relative deviation between simulation and analytic results

4.5.2 Multi-criteria analysis

The choice of an initial stock level in combination with an emergency repair speed determines the cost and service behavior of the model. The optimal solution depends on the parameters of the system (i.e. item unit price, inventory holding cost, penalty cost, normal and emergency repair cost) and on the strategic goal of the management. Next to cost minimization, restrictions with respect to the service performance can be very important. For example, a cost-optimal solution with a fill rate performance of seventy percent may not be acceptable in practice. Analysis of a chosen solution with respect to several performance criteria is therefore crucial. We illustrate this multi-criteria analysis with a numerical example. In table 4.4 the service and cost performance of the ERM for an arbitrary situation ($\lambda=0.01$ and $\rho=0.5$) is presented. The total cost is calculated for two cases. Case 1 represents a situation with relative low penalty cost, high holding cost, and low emergency repair cost. Case 2 represents the opposite situation with relative high penalty cost, low holding cost, and high emergency repair cost. The following performance measures are presented in table 4.4:

- FR : fill rate
- $W^*\mu$: normalized response time for a backorder
- $v(5)$: probability that the response time for a backorder exceeds five days
- $v(10)$: probability that the response time for a backorder exceeds ten days
- TC(1) : total cost for case 1 ($K=1000, h=0.3, p=100, f^n=0.1, f_{max}=0.5$)
- TC(2) : total cost for case 2 ($K=1000, h=0.1, p=1000, f^n=0.1, f_{max}=0.8$)

If the strategic goal of the management is to minimize total cost then the optimal solution is ($S=2, \tau/\mu=10$) for case 1 and ($S=4, \tau/\mu=10$) for case 2. However, if market conditions require a minimum fill rate performance of 95%, then the optimal solution for case 1 is not feasible. In this case the optimal solution for case 1 becomes ($S=3, \tau/\mu=5$). If the service requirements state that the probability of a backorder duration longer than ten days should be less than five percent, the optimal solution becomes ($S=5, \tau/\mu=10$) for both cases. This simple example shows the complex trade-off

between service requirements and cost-optimal solutions. Next to the parameter values of the system, the strategic objectives influence the choice of an initial stock level in combination with an emergency repair rate.

| S | τ/μ | FR | $W^*\mu$ | $v(5)$ | $v(10)$ | TC(1) | TC(2) |
|---|------------|-------|----------|--------|---------|-------|--------|
| 0 | 1 | 0 | 1.000 | 0.94 | 0.87 | 51.00 | 501.00 |
| | 5 | 0 | 0.200 | 0.62 | 0.38 | 12.78 | 104.11 |
| | 10 | 0 | 0.100 | 0.37 | 0.13 | 10.00 | 58.00 |
| 1 | 1 | 0.607 | 0.541 | 0.85 | 0.72 | 12.15 | 107.70 |
| | 5 | 0.663 | 0.166 | 0.56 | 0.31 | 4.95 | 30.26 |
| | 10 | 0.666 | 0.091 | 0.34 | 0.11 | 4.40 | 18.71 |
| 2 | 1 | 0.910 | 0.362 | 0.77 | 0.59 | 3.88 | 17.74 |
| | 5 | 0.922 | 0.143 | 0.51 | 0.25 | 2.96 | 7.26 |
| | 10 | 0.923 | 0.083 | 0.30 | 0.09 | 2.90 | 5.18 |
| 3 | 1 | 0.986 | 0.270 | 0.70 | 0.49 | 3.25 | 3.62 |
| | 5 | 0.987 | 0.125 | 0.46 | 0.20 | 3.16 | 2.54 |
| | 10 | 0.987 | 0.077 | 0.27 | 0.07 | 3.16 | 2.27 |
| 4 | 1 | 0.998 | 0.214 | 0.63 | 0.40 | 3.90 | 2.15 |
| | 5 | 0.998 | 0.111 | 0.41 | 0.17 | 3.89 | 2.05 |
| | 10 | 0.998 | 0.071 | 0.25 | 0.06 | 3.89 | 2.03 |
| 5 | 1 | 1.000 | 0.177 | 0.57 | 0.33 | 4.70 | 2.25 |
| | 5 | 1.000 | 0.100 | 0.37 | 0.14 | 4.70 | 2.24 |
| | 10 | 1.000 | 0.067 | 0.22 | 0.05 | 4.70 | 2.24 |

Table 4.4: Cost and service performance for $\lambda=0.01$ and $\rho=0.5$

4.5.3 Cost behavior and policy evaluation

Given the cost structure defined in Section 4.3, values of the model parameters (demand rate, repair rates, cost factors), and values of the decision variables S (initial stock level) and x (trigger level), we can compute the total cost. In our experiment we vary the following five parameters:

- Daily demand rate (λ);
- Utilization rate ($\rho=\lambda/\mu$);
- Annual inventory holding cost per unit as percentage of the unit price (h);
- Maximum emergency repair cost (with $\tau_{max}=10\mu$) as percentage of the unit price (f_{max});
- Penalty cost per day for a backordered item (p).

Emergency Repair Model

Next to this, we consider three values of the relative emergency repair rate: $\tau/\mu=2$ (table 4.5), $\tau/\mu=5$ (table 4.6) and $\tau/\mu=10$ (table 4.7). The remaining parameters are fixed as follows: $K=1000$ and $f^n=0.1$. For each value of the relative emergency repair rate we examined 36 cases. For each case the minimum cost was calculated by enumeration over the decision variables S and x . The results are presented in tables 4.5, 4.6, and 4.7. The minimum cost is denoted by TC^{opt} , the associated stock level by S^{opt} , and the emergency trigger level by x^{opt} . We also calculated the minimum cost for two alternative policies:

- Policy 1* : Emergency repair is not possible (i.e. $\tau = \mu$).
Policy 2 : Emergency repair is applied when the inventory on hand is zero (i.e. $x = 0$).

The associated total cost and initial stock level are denoted by TC^1 and S^1 (policy 1) and TC^2 and S^2 (policy 2). We also calculated the cost reduction (%) that is obtained when applying the optimal ERM-policy in comparison with policy 1 and policy 2. The results show that significant cost savings can be realized when allowing for fast emergency repair in combination with the use of emergency trigger levels. The results in table 4.8 show that the ERM-policy realizes an overall average cost reduction of 7.2% in comparison with policy 1 and an overall average cost reduction of 3.6% in comparison with policy 2. A maximum cost reduction of 34.5% is realized in comparison with policy 1 and a maximum cost reduction of 30.1% is realized in comparison with policy 2. For both alternative policies the results indicate that the highest cost savings (both average and maximum cost savings) are realized for the intermediate value of the emergency repair rate ($\tau/\mu=5$). For a low emergency repair rate ($\tau/\mu=2$) and a high emergency repair rate ($\tau/\mu=10$) the cost savings are in general lower. The optimal emergency repair speed for a given situation depends on the cost structure and the system parameters. Another observation is that in general the highest cost savings (in comparison with both policies) are obtained for a low value of the demand rate ($\lambda=0.01$). The benefit of using emergency repair (in terms of cost reduction) is apparently maximal in situations with low demand items, which is typical for a service part environment.

Finally, the results show that the ERM-policy is always as least as good as the other two policies. This is logical, since the ERM-policy dominates the other two policies. Table 4.9 shows that in 88.9% of all cases the ERM-policy performs better than policy 1 and in 63.9% of all cases better than policy 2. Especially when the emergency repair rate is relatively low, the ERM-policy outperforms the other policies more often. Comparing policy 1 with policy 2 shows that the latter policy is better in 82.4% of all cases. Using emergency repair with a trigger level equal to zero is in most cases better than using no emergency repair at all, especially for low values of the emergency repair rate. However, this is not always the case, since in 15.7% of all cases policy 1 performs actually better than policy 2.

| λ | ρ | h | f_{max} | p | TC^{opt} (S^{opt}, x^{opt}) | TC^1 | S^1 | % | TC^2 | S^2 | % |
|-----------|--------|-----|-----------|------|-----------------------------------|--------|-------|------|--------|-------|------|
| 0.01 | 0.5 | 0.1 | 0.8 | 100 | 1.84 (3,0) | 1.88 | 3 | 2.1 | 1.84 | 3 | 0.0 |
| | | | | 1000 | 2.06 (4,1) | 2.15 | 4 | 4.2 | 2.10 | 4 | 1.9 |
| | | 0.4 | 0.4 | 100 | 3.41 (2,1) | 3.94 | 3 | 13.5 | 3.75 | 2 | 9.1 |
| | | | | 1000 | 4.33 (3,2) | 5.02 | 4 | 13.7 | 4.98 | 4 | 13.1 |
| | 1.0 | 0.1 | 0.8 | 100 | 2.08 (4,1) | 2.16 | 5 | 3.7 | 2.14 | 4 | 2.8 |
| | | | | 1000 | 2.36 (5,2) | 2.46 | 6 | 4.1 | 2.44 | 6 | 3.3 |
| | | 0.4 | 0.4 | 100 | 4.09 (3,2) | 4.73 | 4 | 13.5 | 4.61 | 4 | 11.3 |
| | | | | 1000 | 5.48 (4,2) | 6.07 | 5 | 9.7 | 5.90 | 5 | 7.1 |
| | 2.0 | 0.1 | 0.8 | 100 | 2.38 (6,2) | 2.51 | 7 | 5.2 | 2.48 | 7 | 4.0 |
| | | | | 1000 | 2.74 (7,3) | 2.94 | 8 | 6.8 | 2.87 | 8 | 4.5 |
| | | 0.4 | 0.4 | 100 | 5.16 (5,2) | 5.98 | 6 | 13.7 | 5.77 | 5 | 10.6 |
| | | | | 1000 | 6.58 (6,3) | 7.87 | 8 | 16.4 | 7.50 | 7 | 12.3 |
| 0.05 | 0.5 | 0.1 | 0.8 | 100 | 5.87 (3,-1) | 5.88 | 3 | 0.2 | 5.88 | 3 | 0.2 |
| | | | | 1000 | 6.11 (4,1) | 6.15 | 4 | 0.7 | 6.11 | 4 | 0.0 |
| | | 0.4 | 0.4 | 100 | 7.86 (2,0) | 7.94 | 3 | 1.0 | 7.86 | 2 | 0.0 |
| | | | | 1000 | 8.72 (3,1) | 9.02 | 4 | 3.3 | 8.98 | 4 | 2.9 |
| | 1.0 | 0.1 | 0.8 | 100 | 6.16 (5,0) | 6.16 | 5 | 0.0 | 6.16 | 5 | 0.0 |
| | | | | 1000 | 6.44 (6,1) | 6.46 | 6 | 0.3 | 6.45 | 6 | 0.2 |
| | | 0.4 | 0.4 | 100 | 8.56 (3,1) | 8.73 | 4 | 1.9 | 8.63 | 4 | 0.8 |
| | | | | 1000 | 9.74 (5,2) | 10.07 | 5 | 3.3 | 9.91 | 5 | 1.7 |
| | 2.0 | 0.1 | 0.8 | 100 | 6.49 (7,0) | 6.51 | 7 | 0.3 | 6.49 | 7 | 0.0 |
| | | | | 1000 | 6.82 (8,2) | 6.94 | 8 | 1.7 | 6.87 | 8 | 0.7 |
| | | 0.4 | 0.4 | 100 | 9.52 (5,1) | 9.98 | 6 | 4.6 | 9.83 | 5 | 3.2 |
| | | | | 1000 | 10.92 (6,3) | 11.87 | 8 | 8.0 | 11.50 | 7 | 5.0 |
| 0.10 | 0.5 | 0.1 | 0.8 | 100 | 10.88 (3,-2) | 10.88 | 3 | 0.0 | 10.93 | 3 | 0.5 |
| | | | | 1000 | 11.12 (4,0) | 11.15 | 4 | 0.3 | 11.12 | 4 | 0.0 |
| | | 0.4 | 0.4 | 100 | 12.93 (3,0) | 12.94 | 3 | 0.1 | 12.93 | 3 | 0.0 |
| | | | | 1000 | 13.86 (3,1) | 14.02 | 4 | 1.1 | 13.99 | 4 | 0.9 |
| | 1.0 | 0.1 | 0.8 | 100 | 11.16 (5,-2) | 11.16 | 5 | 0.0 | 11.17 | 5 | 0.1 |
| | | | | 1000 | 11.45 (6,0) | 11.46 | 6 | 0.1 | 11.45 | 6 | 0.0 |
| | | 0.4 | 0.4 | 100 | 13.66 (4,0) | 13.73 | 4 | 0.5 | 13.66 | 4 | 0.0 |
| | | | | 1000 | 14.80 (5,1) | 15.07 | 5 | 1.8 | 14.91 | 5 | 0.7 |
| | 2.0 | 0.1 | 0.8 | 100 | 11.50 (7,-1) | 11.51 | 7 | 0.1 | 11.50 | 7 | 0.0 |
| | | | | 1000 | 11.84 (8,1) | 11.94 | 8 | 0.8 | 11.88 | 8 | 0.3 |
| | | 0.4 | 0.4 | 100 | 14.86 (6,0) | 14.98 | 6 | 0.8 | 14.86 | 6 | 0.0 |
| | | | | 1000 | 16.13 (7,2) | 16.87 | 8 | 4.4 | 16.51 | 7 | 2.3 |

Table 4.5: Optimal cost results when $\nu\mu=2$ for the situations with trigger level (TC^{opt}), without emergency repair (TC^1) and with trigger level equal to zero (TC^2).

Emergency Repair Model

| λ | ρ | h | f_{max} | p | TC^{opt} (S^{opt}, x^{opt}) | TC^1 | S^1 | % | TC^2 | S^2 | % |
|-----------|--------|-----|-----------|------|-----------------------------------|--------|-------|------|--------|-------|------|
| 0.01 | 0.5 | 0.1 | 0.8 | 100 | 1.81 (3,0) | 1.88 | 3 | 5.7 | 1.81 | 3 | 0.0 |
| | | | | 1000 | 2.03 (4,1) | 2.15 | 4 | 12.2 | 2.05 | 4 | 8.9 |
| | | 0.4 | 0.4 | 100 | 3.35 (2,0) | 3.94 | 3 | 19.1 | 3.35 | 2 | 1.8 |
| | | | | 1000 | 4.10 (3,1) | 5.02 | 4 | 22.9 | 4.57 | 3 | 10.5 |
| | 1.0 | 0.1 | 0.8 | 100 | 2.06 (4,0) | 2.16 | 5 | 7.6 | 2.06 | 4 | 2.6 |
| | | | | 1000 | 2.27 (5,1) | 2.46 | 6 | 12.7 | 2.42 | 6 | 6.8 |
| | | 0.4 | 0.4 | 100 | 3.88 (3,1) | 4.73 | 4 | 27.2 | 4.16 | 3 | 13.1 |
| | | | | 1000 | 4.87 (4,2) | 6.07 | 5 | 27.5 | 5.71 | 5 | 23.1 |
| | 2.0 | 0.1 | 0.8 | 100 | 2.33 (6,1) | 2.51 | 7 | 12.7 | 2.39 | 6 | 5.3 |
| | | | | 1000 | 2.60 (7,2) | 2.94 | 8 | 17.6 | 2.79 | 8 | 12.1 |
| | | 0.4 | 0.4 | 100 | 4.63 (3,2) | 5.98 | 6 | 34.5 | 5.22 | 5 | 23.7 |
| | | | | 1000 | 5.68 (4,3) | 7.87 | 8 | 34.5 | 7.13 | 7 | 30.1 |
| 0.05 | 0.5 | 0.1 | 0.8 | 100 | 5.88 (3,-2) | 5.88 | 3 | 0.0 | 5.97 | 3 | 1.2 |
| | | | | 1000 | 6.07 (4,0) | 6.15 | 4 | 0.8 | 6.07 | 4 | 0.0 |
| | | 0.4 | 0.4 | 100 | 7.77 (2,0) | 7.94 | 3 | 5.6 | 7.77 | 2 | 0.0 |
| | | | | 1000 | 8.52 (3,1) | 9.02 | 4 | 9.4 | 8.64 | 3 | 1.2 |
| | 1.0 | 0.1 | 0.8 | 100 | 6.16 (5,-1) | 6.16 | 5 | 0.4 | 6.18 | 5 | 0.4 |
| | | | | 1000 | 6.43 (6,0) | 6.46 | 6 | 3.3 | 6.43 | 6 | 0.0 |
| | | 0.4 | 0.4 | 100 | 8.50 (3,0) | 8.73 | 4 | 7.1 | 8.50 | 3 | 0.0 |
| | | | | 1000 | 9.37 (4,1) | 10.07 | 5 | 9.8 | 9.73 | 5 | 6.8 |
| | 2.0 | 0.1 | 0.8 | 100 | 6.49 (7,-1) | 6.51 | 7 | 1.4 | 6.49 | 7 | 0.0 |
| | | | | 1000 | 6.77 (8,1) | 6.94 | 8 | 5.6 | 6.81 | 8 | 1.9 |
| | | 0.4 | 0.4 | 100 | 9.43 (5,0) | 9.98 | 6 | 9.3 | 9.43 | 5 | 1.2 |
| | | | | 1000 | 10.53 (6,2) | 11.87 | 8 | 12.9 | 11.15 | 7 | 8.7 |
| 0.10 | 0.5 | 0.1 | 0.8 | 100 | 10.88 (3,-4) | 10.88 | 3 | 0.0 | 11.02 | 4 | 2.3 |
| | | | | 1000 | 11.10 (4,0) | 11.15 | 4 | 0.4 | 11.10 | 4 | 0.0 |
| | | 0.4 | 0.4 | 100 | 12.93 (3,-1) | 12.94 | 3 | 0.6 | 13.00 | 3 | 0.5 |
| | | | | 1000 | 13.73 (3,0) | 14.02 | 4 | 5.2 | 13.73 | 3 | 0.0 |
| | 1.0 | 0.1 | 0.8 | 100 | 11.17 (5,-2) | 11.16 | 5 | 0.0 | 11.23 | 5 | 1.8 |
| | | | | 1000 | 11.44 (6,0) | 11.46 | 6 | 1.8 | 11.44 | 6 | 0.0 |
| | | 0.4 | 0.4 | 100 | 13.69 (4,-1) | 13.73 | 4 | 2.1 | 13.70 | 4 | 0.0 |
| | | | | 1000 | 14.73 (5,1) | 15.07 | 5 | 4.6 | 14.75 | 5 | 2.9 |
| | 2.0 | 0.1 | 0.8 | 100 | 11.51 (7,-2) | 11.51 | 7 | 0.3 | 11.55 | 7 | 0.7 |
| | | | | 1000 | 11.82 (8,0) | 11.94 | 8 | 2.6 | 11.82 | 8 | 0.4 |
| | | 0.4 | 0.4 | 100 | 14.70 (5,0) | 14.98 | 6 | 4.5 | 14.83 | 5 | 0.0 |
| | | | | 1000 | 15.95 (6,1) | 16.87 | 8 | 7.1 | 16.17 | 7 | 3.8 |

Table 4.6: Optimal cost results when $\nu\mu=5$ for the situations with trigger level (TC^{opt}), without emergency repair (TC^1) and with trigger level equal to zero (TC^2).

| λ | ρ | h | f_{max} | p | TC^{opt} (S^{opt}, x^{opt}) | TC^1 | S^1 | % | TC^2 | S^2 | % |
|-----------|--------|-----|-----------|------|-----------------------------------|--------|-------|------|--------|-------|------|
| 0.01 | 0.5 | 0.1 | 0.8 | 100 | 1.83 (3,0) | 1.88 | 3 | 4.0 | 1.83 | 3 | 0.0 |
| | | | | 1000 | 2.03 (4,0) | 2.15 | 4 | 9.9 | 2.03 | 4 | 0.0 |
| | | 0.4 | 0.4 | 100 | 3.24 (2,0) | 3.94 | 3 | 19.5 | 3.24 | 2 | 0.0 |
| | | | | 1000 | 4.08 (3,1) | 5.02 | 4 | 27.1 | 4.28 | 3 | 11.1 |
| | 1.0 | 0.1 | 0.8 | 100 | 2.05 (4,0) | 2.16 | 5 | 7.0 | 2.05 | 4 | 0.0 |
| | | | | 1000 | 2.25 (5,1) | 2.46 | 6 | 15.2 | 2.33 | 5 | 7.7 |
| | | 0.4 | 0.4 | 100 | 3.94 (3,0) | 4.73 | 4 | 21.8 | 3.94 | 3 | 3.0 |
| | | | | 1000 | 4.74 (3,1) | 6.07 | 5 | 27.6 | 5.47 | 4 | 16.7 |
| | 2.0 | 0.1 | 0.8 | 100 | 2.35 (6,0) | 2.51 | 7 | 10.5 | 2.35 | 6 | 0.0 |
| | | | | 1000 | 2.58 (7,1) | 2.94 | 8 | 17.3 | 2.75 | 8 | 9.0 |
| | | 0.4 | 0.4 | 100 | 4.53 (4,1) | 5.98 | 6 | 31.7 | 4.99 | 5 | 13.0 |
| | | | | 1000 | 5.51 (5,2) | 7.87 | 8 | 33.9 | 6.92 | 7 | 24.9 |
| 0.05 | 0.5 | 0.1 | 0.8 | 100 | 5.88 (3,-4) | 5.88 | 3 | 0.0 | 6.02 | 4 | 4.0 |
| | | | | 1000 | 6.07 (4,0) | 6.15 | 4 | 1.0 | 6.07 | 4 | 0.0 |
| | | 0.4 | 0.4 | 100 | 7.93 (3,-1) | 7.94 | 3 | 1.4 | 7.99 | 3 | 0.0 |
| | | | | 1000 | 8.43 (3,0) | 9.02 | 4 | 10.2 | 8.43 | 3 | 0.0 |
| | 1.0 | 0.1 | 0.8 | 100 | 6.16 (5,-3) | 6.16 | 5 | 0.0 | 6.23 | 5 | 2.6 |
| | | | | 1000 | 6.41 (5,0) | 6.46 | 6 | 3.9 | 6.41 | 5 | 0.0 |
| | | 0.4 | 0.4 | 100 | 8.65 (4,0) | 8.73 | 4 | 5.3 | 8.65 | 4 | 0.0 |
| | | | | 1000 | 9.51 (4,1) | 10.07 | 5 | 8.7 | 9.64 | 5 | 1.3 |
| | 2.0 | 0.1 | 0.8 | 100 | 6.50 (7,-1) | 6.51 | 7 | 0.7 | 6.54 | 7 | 0.7 |
| | | | | 1000 | 6.78 (8,0) | 6.94 | 8 | 5.0 | 6.78 | 8 | 0.0 |
| | | 0.4 | 0.4 | 100 | 9.44 (5,0) | 9.98 | 6 | 8.2 | 9.44 | 5 | 0.0 |
| | | | | 1000 | 10.41 (6,1) | 11.87 | 8 | 12.9 | 10.96 | 7 | 5.3 |
| 0.10 | 0.5 | 0.1 | 0.8 | 100 | 10.88 (3,-4) | 10.88 | 3 | 0.0 | 11.08 | 4 | 5.7 |
| | | | | 1000 | 11.13 (4,0) | 11.15 | 4 | 0.1 | 11.13 | 4 | 0.0 |
| | | 0.4 | 0.4 | 100 | 12.94 (3,-4) | 12.94 | 3 | 0.0 | 13.18 | 3 | 2.9 |
| | | | | 1000 | 13.62 (3,0) | 14.02 | 4 | 5.9 | 13.62 | 3 | 0.0 |
| | 1.0 | 0.1 | 0.8 | 100 | 11.17 (5,-4) | 11.16 | 5 | 0.0 | 11.34 | 5 | 4.5 |
| | | | | 1000 | 11.44 (6,0) | 11.46 | 6 | 1.8 | 11.44 | 6 | 0.0 |
| | | 0.4 | 0.4 | 100 | 13.72 (4,-2) | 13.73 | 4 | 0.1 | 13.89 | 4 | 1.1 |
| | | | | 1000 | 14.69 (5,0) | 15.07 | 5 | 4.1 | 14.69 | 5 | 0.0 |
| | 2.0 | 0.1 | 0.8 | 100 | 11.51 (7,-4) | 11.51 | 7 | 0.0 | 11.66 | 7 | 2.9 |
| | | | | 1000 | 11.81 (8,0) | 11.94 | 8 | 2.7 | 11.81 | 8 | 0.0 |
| | | 0.4 | 0.4 | 100 | 14.90 (6,-1) | 14.98 | 6 | 2.8 | 14.94 | 6 | 0.0 |
| | | | | 1000 | 15.97 (6,1) | 16.87 | 8 | 7.0 | 16.02 | 7 | 1.9 |

Table 4.7: Optimal cost results when $\nu\mu=10$ for the situations with trigger level (TC^{opt}), without emergency repair (TC^1) and with trigger level equal to zero (TC^2).

| τ/μ | Policy 1 | | | Policy 2 | | |
|------------|------------|------------|----------------|------------|------------|----------------|
| | <i>min</i> | <i>max</i> | <i>average</i> | <i>min</i> | <i>max</i> | <i>average</i> |
| 2 | 0.0 | 16.4 | 3.9 | 0.0 | 13.1 | 2.8 |
| 5 | 0.0 | 34.5 | 9.1 | 0.0 | 30.1 | 4.8 |
| 10 | 0.0 | 33.9 | 8.5 | 0.0 | 24.9 | 3.3 |
| overall | 0.0 | 34.5 | 7.2 | 0.0 | 30.1 | 3.6 |

Table 4.8: Minimum, maximum, and average cost savings (%) of the ERM-policy in comparison with policy 1 and policy 2.

| τ/μ | $TC^{opt} < TC^1$ | $TC^{opt} < TC^2$ | $TC^1 < TC^2$ |
|------------|-------------------|-------------------|---------------|
| 2 | 91.7 % | 69.4 % | 88.9 % |
| 5 | 91.7 % | 69.4 % | 83.3 % |
| 10 | 83.3 % | 52.8 % | 75.0 % |
| overall | 88.9 % | 63.9 % | 82.4 % |

Table 4.9: Cost performance evaluation of the different policies (expressed in percentage of all cases).

4.5.4 Sensitivity analysis

The simulation results presented in Section 4.5.1 show that the service performance of the Emergency Repair Model, especially the fill rate performance, is almost insensitive with respect to the choice of repair lead time distribution. The average deviation between the fill rate obtained from simulation with deterministic repair lead times, and the fill rate obtained from the analytic model, is only 0.1 percent (see table 4.3). In this section we investigate if this strong insensitivity result with respect to the fill rate performance holds for other repair lead time distributions as well. Furthermore, we calculate the fill rate performance of the model when assuming an Erlang distribution instead of an exponential distribution for the inter arrival times of failed parts.

In tables 4.10 and 4.11 we present the simulation results for different distributions of the repair lead times and the inter arrival times of failed service parts. We used the same data set as before (see tables 4.1 and 4.2). In practice, the normal repair process is often represented by a repair shop in which repair jobs can be delayed substantially. The lognormal distribution is an example of a

skewed distribution with a thick tail that represents such a process. We therefore used the lognormal distribution for the normal repair process in combination with exponentially distributed emergency repair lead times (in accordance with the model assumptions) and deterministic emergency repair lead times (in accordance with the realistic situation of subcontracting repair work). The simulation results from Section 4.5.1 (assuming deterministic repair lead times for both repair processes, see tables 4.1 and 4.2) are presented as well. With respect to the arrival process of failed service parts, we used the Erlang-2 and Erlang-10 distribution for the inter arrival times of failed parts. These distributions represent a more stable arrival process of failed parts. All distributions are chosen such that the mean of the distributions is the same in all situations. The following notation is used in tables 4.10 and 4.11:

ERM = Emergency Repair Model.

E,L,E = Exponentially distributed inter arrival times, lognormal distributed normal repair lead times, and exponentially distributed emergency repair lead times.

E,L,D = Exponentially distributed inter arrival times, lognormal distributed normal repair lead times, and deterministic emergency repair lead times.

E,D,D = Exponentially distributed inter arrival times, deterministic normal repair lead times, and deterministic emergency repair lead times.

E_k ,E,E = Erlang- k ($k=2,10$) distributed inter arrival times, exponentially distributed normal repair lead times, and exponentially distributed emergency repair lead times.

The results indicate that the fill rate performance is insensitive with respect to the choice of repair lead time distribution. Whether the repair lead times are exponentially distributed, lognormal distributed, or deterministic, the differences in fill rate performance are negligible. Even the use of the lognormal distribution for normal repair lead times does not affect the fill rate performance significantly. The average deviation in fill rate performance when assuming lognormal distributed normal repair lead times and exponentially distributed emergency repair lead times (i.e. E,L,E-cases in tables 4.10 and 4.11) is only 0.1 percent. The average deviation in fill rate performance when assuming lognormal distributed normal repair lead times and deterministic emergency repair lead times (i.e. E,L,D-cases in tables 4.10 and 4.11) is only 0.4 percent. Note that there is only a significant difference in fill rate performance between the E,L,E-cases and the E,L,D-cases when the fill rates are low ($<60\%$) and $\tau=\mu$. The fill rate performance when assuming deterministic emergency repairs (i.e. with no variation) is in these situations slightly better, since emergency repair is often applied. In general, we can conclude that the assumption of exponentially distributed repair lead times in the Emergency Repair Model is not very restrictive with respect to the applicability of our model. Moreover, the fill rate performance of the Emergency Repair Model is to a large extent independent of the choice of the (normal and emergency) repair lead time distribution.

Emergency Repair Model

In our model we assume that the arrival process of failed parts at the inventory location is represented by a Poisson process, i.e. the inter arrival times of failed parts are exponentially distributed. The results in tables 4.10 and 4.11 show that this assumption is very important. The fill rate performance, obtained from simulation with Erlang-2 distributed inter arrival times of failed service part (i.e. E_2, E, E -cases in tables 4.10 and 4.11), increases dramatically. The fill rate performance increases even further when assuming Erlang-10 distributed inter arrival times (i.e. E_{10}, E, E -cases in tables 4.10 and 4.11). From these results we can conclude that the assumption of Poisson arrivals of failed parts is crucial with respect to the applicability of the model. A similar observation is made by Smith and Dekker (1996) for other models. The Emergency Repair Model is only applicable in those situations in which the arrival process of failed parts can be represented by a Poisson process.

| ρ | S | τ/μ | ERM | E, L, E | E, L, D | E, D, D | E_2, E, E | E_{10}, E, E |
|--------|---|------------|-------|-----------|-----------|-----------|-------------|----------------|
| 0.1 | 1 | 1 | 0.905 | 0.904 | 0.905 | 0.905 | 0.972 | 0.999 |
| | | 5 | 0.909 | 0.909 | 0.909 | 0.909 | 0.972 | 0.999 |
| | | 10 | 0.909 | 0.909 | 0.909 | 0.909 | 0.972 | 0.999 |
| | 2 | 1 | 0.995 | 0.995 | 0.995 | 0.995 | 1.000 | 1.000 |
| | | 5 | 0.995 | 0.996 | 0.995 | 0.995 | 1.000 | 1.000 |
| | | 10 | 0.995 | 0.995 | 0.995 | 0.995 | 1.000 | 1.000 |
| | 3 | 1 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | 5 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | 10 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.5 | 1 | 1 | 0.607 | 0.604 | 0.617 | 0.607 | 0.706 | 0.815 |
| | | 5 | 0.663 | 0.663 | 0.664 | 0.665 | 0.749 | 0.838 |
| | | 10 | 0.666 | 0.665 | 0.667 | 0.667 | 0.750 | 0.839 |
| | 2 | 1 | 0.910 | 0.910 | 0.910 | 0.910 | 0.964 | 0.993 |
| | | 5 | 0.922 | 0.921 | 0.922 | 0.922 | 0.968 | 0.994 |
| | | 10 | 0.923 | 0.922 | 0.923 | 0.923 | 0.968 | 0.994 |
| | 3 | 1 | 0.986 | 0.986 | 0.986 | 0.986 | 0.998 | 1.000 |
| | | 5 | 0.987 | 0.987 | 0.987 | 0.987 | 0.998 | 1.000 |
| | | 10 | 0.987 | 0.987 | 0.987 | 0.987 | 0.998 | 1.000 |
| | 4 | 1 | 0.998 | 0.998 | 0.998 | 0.998 | 1.000 | 1.000 |
| | | 5 | 0.998 | 0.998 | 0.998 | 0.998 | 1.000 | 1.000 |
| | | 10 | 0.998 | 0.998 | 0.998 | 0.998 | 1.000 | 1.000 |
| | 5 | 1 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | 5 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| | | 10 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

Table 4.10: Fill rate performance for different distributions of the repair processes and arrival process of failed service parts for $\rho=0.1$ and $\rho=0.5$.

| ρ | S | τ/μ | ERM | E ₁ L,E | E ₁ L,D | E ₁ D,D | E ₂ ,E,E | E ₁₀ ,E,E |
|--------|---|------------|-------|--------------------|--------------------|--------------------|---------------------|----------------------|
| 1 | 1 | 1 | 0.368 | 0.363 | 0.385 | 0.369 | 0.422 | 0.484 |
| | | 5 | 0.491 | 0.492 | 0.496 | 0.495 | 0.549 | 0.612 |
| | | 10 | 0.498 | 0.498 | 0.499 | 0.499 | 0.554 | 0.616 |
| | 2 | 1 | 0.736 | 0.734 | 0.738 | 0.736 | 0.818 | 0.897 |
| | | 5 | 0.794 | 0.794 | 0.796 | 0.796 | 0.858 | 0.920 |
| | | 10 | 0.798 | 0.798 | 0.799 | 0.799 | 0.862 | 0.919 |
| | 3 | 1 | 0.920 | 0.920 | 0.918 | 0.920 | 0.965 | 0.991 |
| | | 5 | 0.935 | 0.935 | 0.936 | 0.935 | 0.971 | 0.993 |
| | | 10 | 0.937 | 0.937 | 0.938 | 0.937 | 0.972 | 0.993 |
| | 4 | 1 | 0.981 | 0.981 | 0.980 | 0.981 | 0.995 | 1.000 |
| | | 5 | 0.984 | 0.984 | 0.984 | 0.984 | 0.996 | 1.000 |
| | | 10 | 0.984 | 0.985 | 0.985 | 0.984 | 0.996 | 1.000 |
| | 5 | 1 | 0.996 | 0.996 | 0.996 | 0.996 | 1.000 | 1.000 |
| | | 5 | 0.997 | 0.997 | 0.997 | 0.997 | 1.000 | 1.000 |
| | | 10 | 0.997 | 0.997 | 0.997 | 0.997 | 1.000 | 1.000 |
| | 6 | 1 | 0.999 | 0.999 | 0.999 | 0.999 | 1.000 | 1.000 |
| | | 5 | 0.999 | 0.999 | 0.999 | 0.999 | 1.000 | 1.000 |
| | | 10 | 0.999 | 1.000 | 0.999 | 0.999 | 1.000 | 1.000 |
| 2 | 1 | 1 | 0.135 | 0.131 | 0.150 | 0.136 | 0.136 | 0.137 |
| | | 5 | 0.318 | 0.317 | 0.323 | 0.324 | 0.346 | 0.376 |
| | | 10 | 0.329 | 0.329 | 0.332 | 0.331 | 0.358 | 0.384 |
| | 2 | 1 | 0.406 | 0.400 | 0.424 | 0.407 | 0.450 | 0.495 |
| | | 5 | 0.583 | 0.582 | 0.588 | 0.588 | 0.633 | 0.686 |
| | | 10 | 0.595 | 0.596 | 0.598 | 0.597 | 0.644 | 0.692 |
| | 3 | 1 | 0.677 | 0.675 | 0.679 | 0.677 | 0.748 | 0.822 |
| | | 5 | 0.777 | 0.777 | 0.780 | 0.779 | 0.832 | 0.885 |
| | | 10 | 0.785 | 0.786 | 0.787 | 0.787 | 0.838 | 0.888 |
| | 4 | 1 | 0.857 | 0.857 | 0.853 | 0.857 | 0.916 | 0.962 |
| | | 5 | 0.898 | 0.899 | 0.899 | 0.899 | 0.940 | 0.973 |
| | | 10 | 0.902 | 0.902 | 0.903 | 0.903 | 0.943 | 0.973 |
| | 5 | 1 | 0.947 | 0.948 | 0.944 | 0.947 | 0.979 | 0.995 |
| | | 5 | 0.960 | 0.960 | 0.960 | 0.960 | 0.984 | 0.996 |
| | | 10 | 0.962 | 0.962 | 0.963 | 0.962 | 0.984 | 0.996 |
| | 6 | 1 | 0.983 | 0.983 | 0.982 | 0.983 | 0.996 | 1.000 |
| | | 5 | 0.987 | 0.987 | 0.987 | 0.987 | 0.996 | 1.000 |
| | | 10 | 0.987 | 0.988 | 0.988 | 0.988 | 0.997 | 1.000 |

Table 4.11: Fill rate performance for different distributions of the repair processes and arrival process of failed service parts for $\rho=1$ and $\rho=2$.

4.6 Comparison with the Muckstadt-Thomas model

The approximate two-echelon model developed by Muckstadt and Thomas (1980) is an extension of the METRIC model in that it allows for fast direct deliveries from the central warehouse (or even from the factory) when a demand at a local warehouse can not be met from stock on hand. It is assumed in this model that an emergency resupply order (i.e. direct delivery) that is triggered to fill a local demand, is always used for this purpose, even if a normal replenishment order arrives earlier at the local warehouse. Furthermore, in this model the trigger level for emergency supply is always equal to zero (i.e. emergency supply is only triggered when a demand occurs in a stock-out situation at a local warehouse).

In our ERM we fill a backordered demand with the first available order that arrives at the warehouse, either normal replenishment or emergency. In addition, the trigger level for emergency repair is in most cases not equal to zero for the cost-optimal solution (see tables 4.5, 4.6, and 4.7). By comparing the Muckstadt-Thomas model with our Emergency Repair Model we can quantify the cost savings and service improvements that are obtained by the following two flexibilities:

- 1) *Filling backorders by normal repair orders that arrive earlier than emergency repair orders.*
- 2) *Allowing non-zero trigger levels for emergency repair.*

In order to compare the two-echelon Muckstadt-Thomas model (MT) with our single-echelon Emergency Repair Model (ERM), we consider one local warehouse and assume infinite supply at the central warehouse. The results are presented in table 4.12 for the situation with $\tau/\mu=2$. The same 36 cases are examined as in table 4.5. For both models the minimum total cost (TC) and the associated service measures (FR and $W*\mu$) are presented. The minimum total cost in the MT-model is obtained by enumeration over the initial stock level S . We also calculated the cost reduction percentage (%) that is realized by using the ERM-model instead of the MT-model. In all cases we see that a significant cost reduction is realized with an average of 7.6% and a maximum of 22.0%. This is caused by the fact that the expected backorder duration in the MT-model is significantly longer than in the ERM. The normalized backorder duration in the MT-model is 50% for all cases since backorders are always satisfied through emergency orders that are twice as fast. As a result the penalty cost in the MT-model is much higher than in the ERM.

The added value of using normal repair orders to fill backordered demand can be evaluated by comparing the results of the MT-model with the results of policy 2 in table 4.5. Policy 2 also prescribes a zero trigger level for emergency repair (as is the case in the MT-model) but allows the use of normal repair (or replenishment) orders to fill backordered demand. Policy 2 realizes an average cost reduction of 5.0% (with a maximum of 12.6%) in comparison with the MT-model.

| λ | ρ | h | f_{max} | p | ERM | | | MT | | | | % |
|-----------|--------|-----|-----------|------|-------|-------|---------|-------|---|-------|---------|------|
| | | | | | TC | FR | $W*\mu$ | TC | S | FR | $W*\mu$ | |
| 0.01 | 0.5 | 0.1 | 0.8 | 100 | 1.84 | 0.986 | 0.204 | 2.00 | 4 | 0.998 | 0.500 | 8.0 |
| | | | | 1000 | 2.06 | 0.999 | 0.146 | 2.27 | 5 | 1.000 | 0.500 | 9.3 |
| | | 0.4 | 0.4 | 100 | 3.41 | 0.946 | 0.205 | 4.06 | 3 | 0.987 | 0.500 | 16.0 |
| | | | | 1000 | 4.33 | 0.995 | 0.147 | 5.23 | 4 | 0.998 | 0.500 | 17.2 |
| | 1.0 | 0.1 | 0.8 | 100 | 2.08 | 0.987 | 0.150 | 2.25 | 5 | 0.997 | 0.500 | 7.6 |
| | | | | 1000 | 2.36 | 0.998 | 0.116 | 2.63 | 6 | 0.999 | 0.500 | 10.3 |
| | | 0.4 | 0.4 | 100 | 4.09 | 0.972 | 0.151 | 5.07 | 4 | 0.985 | 0.500 | 19.3 |
| | | | | 1000 | 5.48 | 0.992 | 0.131 | 6.74 | 6 | 0.999 | 0.500 | 18.7 |
| | 2.0 | 0.1 | 0.8 | 100 | 2.38 | 0.993 | 0.109 | 2.72 | 7 | 0.997 | 0.500 | 12.5 |
| | | | | 1000 | 2.74 | 0.999 | 0.090 | 3.11 | 9 | 1.000 | 0.500 | 11.9 |
| | | 0.4 | 0.4 | 100 | 5.16 | 0.979 | 0.121 | 6.60 | 6 | 0.988 | 0.500 | 21.8 |
| | | | | 1000 | 6.58 | 0.996 | 0.099 | 8.44 | 8 | 0.999 | 0.500 | 22.0 |
| 0.05 | 0.5 | 0.1 | 0.8 | 100 | 5.87 | 0.986 | 0.256 | 6.00 | 4 | 0.998 | 0.500 | 2.2 |
| | | | | 1000 | 6.11 | 0.999 | 0.146 | 6.27 | 5 | 1.000 | 0.500 | 2.6 |
| | | 0.4 | 0.4 | 100 | 7.86 | 0.918 | 0.254 | 8.08 | 3 | 0.987 | 0.500 | 2.7 |
| | | | | 1000 | 8.72 | 0.990 | 0.171 | 9.23 | 4 | 0.998 | 0.500 | 5.5 |
| | 1.0 | 0.1 | 0.8 | 100 | 6.16 | 0.997 | 0.150 | 6.26 | 5 | 0.997 | 0.500 | 1.6 |
| | | | | 1000 | 6.44 | 1.000 | 0.116 | 6.63 | 6 | 0.999 | 0.500 | 2.9 |
| | | 0.4 | 0.4 | 100 | 8.56 | 0.951 | 0.175 | 9.09 | 4 | 0.985 | 0.500 | 5.8 |
| | | | | 1000 | 9.74 | 0.998 | 0.116 | 10.74 | 6 | 0.999 | 0.500 | 9.3 |
| | 2.0 | 0.1 | 0.8 | 100 | 6.49 | 0.996 | 0.123 | 6.73 | 7 | 0.997 | 0.500 | 3.6 |
| | | | | 1000 | 6.82 | 0.999 | 0.090 | 7.11 | 9 | 1.000 | 0.500 | 4.1 |
| | | 0.4 | 0.4 | 100 | 9.52 | 0.967 | 0.137 | 10.62 | 6 | 0.988 | 0.500 | 10.4 |
| | | | | 1000 | 10.92 | 0.996 | 0.099 | 12.44 | 8 | 0.999 | 0.500 | 12.2 |
| 0.10 | 0.5 | 0.1 | 0.8 | 100 | 10.88 | 0.986 | 0.268 | 11.01 | 4 | 0.998 | 0.500 | 1.2 |
| | | | | 1000 | 11.12 | 0.998 | 0.171 | 11.27 | 5 | 1.000 | 0.500 | 1.3 |
| | | 0.4 | 0.4 | 100 | 12.93 | 0.986 | 0.204 | 13.10 | 3 | 0.987 | 0.500 | 1.3 |
| | | | | 1000 | 13.86 | 0.990 | 0.171 | 14.24 | 4 | 0.998 | 0.500 | 2.7 |
| | 1.0 | 0.1 | 0.8 | 100 | 11.16 | 0.996 | 0.186 | 11.27 | 5 | 0.997 | 0.500 | 1.0 |
| | | | | 1000 | 11.45 | 0.999 | 0.132 | 11.63 | 6 | 0.999 | 0.500 | 1.5 |
| | | 0.4 | 0.4 | 100 | 13.66 | 0.983 | 0.175 | 14.11 | 4 | 0.985 | 0.500 | 3.2 |
| | | | | 1000 | 14.80 | 0.997 | 0.131 | 15.74 | 6 | 0.999 | 0.500 | 6.0 |
| | 2.0 | 0.1 | 0.8 | 100 | 11.50 | 0.996 | 0.141 | 11.74 | 8 | 0.999 | 0.500 | 2.0 |
| | | | | 1000 | 11.84 | 0.999 | 0.099 | 12.11 | 9 | 1.000 | 0.500 | 2.2 |
| | | 0.4 | 0.4 | 100 | 14.86 | 0.985 | 0.138 | 15.64 | 6 | 0.988 | 0.500 | 5.0 |
| | | | | 1000 | 16.13 | 0.998 | 0.099 | 17.44 | 8 | 0.999 | 0.500 | 7.5 |

Table 4.12: Performance evaluation of the Muckstadt-Thomas model (MT) and the ERM-model with $\tau/\mu=2$

Emergency Repair Model

From this we can conclude that in general two-third of the cost savings can be attributed to the use of normal repair orders filling backordered demand. The extra cost savings that can be realized by also using non-zero trigger levels for emergency repair is approximately one-third of the total cost in the MT-model.

4.7 Conclusions

In this chapter we investigated the use of repair flexibility by developing the Emergency Repair Model. In contrast with most other models in this field, the Emergency Repair Model presents an *exact* analysis. The use of fast emergency repair and the conditions under which to use this flexibility were examined. We can summarize the main conclusions as follows:

- 1) The criterion for applying emergency repair flexibility, i.e. the emergency trigger level, is an important decision variable in the trade-off problem. The numerical results show that, depending on the parameter setting, negative, positive and zero trigger levels can be cost-optimal. This is an important conclusion because in practice one often only uses emergency repair flexibility when the physical stock on hand drops to zero (i.e. trigger level equal to zero). The numerical results indicate that even in many of these situations significant cost savings can be realized. However, there are situations in which using no emergency repair is better than using emergency repair with a zero trigger level.
- 2) The cost structure is very important for the trade-off problem. In our analysis we assumed a cost structure that captures a number of relevant cost factors, such as inventory holding cost, repair cost, and penalty cost for backorders. Other cost structures can be used as well, depending on the practical circumstances. We used the initial stock level and the emergency trigger level as decision variables when minimizing the total cost function. It is also possible to use the emergency repair speed as a decision variable in the cost minimization procedure. The 'optimal' solution is determined by the strategic goal. Requirements with respect to service performance (e.g. in terms of fill rate or backorder duration) can result in a solution that is not cost-minimal.
- 3) We proved for two extreme situations (no emergency repair and instantaneous emergency repair) that the performance of the model only depends on the mean of the repair time distribution and not on the repair time distribution itself. Simulation results indicate that this strong insensitivity result holds as well for intermediate values of the emergency repair rate. An explanation for this insensitivity result with respect to the repair time distribution is the relatively long inter-arrival times of failed parts at the repair shop. As a consequence, variations in the repair times have a negligible effect on the performance characteristics.

- 4) Although the performance of the model showed a high degree of insensitivity with respect to the choice of repair lead time distribution, the assumption of Poisson arrivals of failed service parts appears to be crucial. Assuming Erlang- k ($k=2,10$) distributed inter arrival times of failed parts instead of exponentially distributed inter arrival times, affects the fill rate performance of the model dramatically. The fill rates increase enormously when the variability of the arrival process decreases. This is an important conclusion with respect to the applicability of models in general that assume Poisson demand. If demand for service parts is more predictable, for example when using preventive maintenance policies for replacing service parts, the assumption of Poisson demand is often not realistic, and hence these models should be applied with care.
- 5) Comparison of the ERM policy with three other policies reveals that significant cost savings can be obtained. From the comparison with policy 1 (no emergency repair) we can conclude that applying emergency repair flexibility can be very beneficial. From the comparison with policy 2 (trigger level equal to zero) we can conclude that allowing for non-zero trigger levels can be very beneficial. Finally, from the comparison with the Muckstadt-Thomas model we can conclude that allowing to fill backorders by normal repair orders that arrive earlier, can also be very beneficial.

In the analysis of the Emergency Repair Model we assume exponentially distributed repair times. Consequently, we cannot use pipeline flexibility in case of a stock-out situation. That is, we cannot check the expected arrival time of outstanding repair orders that might arrive earlier than a newly issued emergency repair order. The added value of using pipeline flexibility is investigated in the simulation study that is presented in Chapter 6. In practice, companies often have an extensive multi-echelon service part supply system in order to realize a high service performance to customers spread over a large geographical area. In the next chapter we investigate the use of supply flexibility in such an inventory system.

Chapter 5

The Emergency Supply Model

5.1 Introduction

The Emergency Repair Model introduced in the previous chapter can be used to make a trade-off between applying emergency procedures and investing in inventory for a *single-echelon* inventory location. In practice most companies use an extensive *multi-echelon* distribution network to supply service parts to customers. The reason that companies use multi-layered distribution systems is the scattering of customers over a large geographical area. The responsibility for supplying service parts to customers in, for example, Europe, often resides with one central service organization. In order to guarantee a high service performance to customers in the north of Sweden as well as to customers in the south of Spain necessitates the use of national and/or regional warehouses. The investment in inventory in such multi-echelon distribution systems is usually very high and therefore it is interesting to consider the implementation of flexibility in these kind of systems. In this chapter we address the issue of supply flexibility in these systems as identified in the SPSS framework of Chapter 3 (see figure 5.1). Practical examples of supply flexibility options are the use of emergency lateral transshipments (pooling flexibility) and direct deliveries (direct shipment flexibility). Lateral

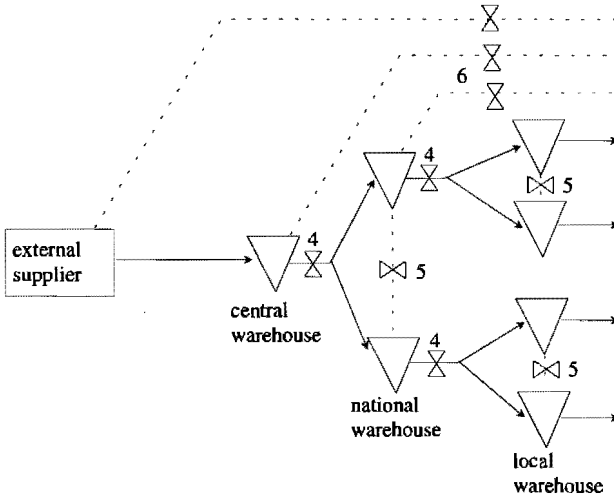


Figure 5.1: Allocation flexibility (4), pooling flexibility (5), and direct shipment flexibility in the SPSS

transhipments are used to fill a demand at a local warehouse that is out of stock from any other local warehouse in the same geographical region that does have excess stock on hand. Direct deliveries are used to fill such a demand from a higher level in the system, e.g., a central warehouse or a production plant.

In this chapter we present a two-echelon inventory model, the so-called Emergency Supply Model (ESM), in which we implicitly address three supply flexibility options: *allocation flexibility*, *pooling flexibility*, and *direct shipment flexibility* (see Alfredsson and Verrijdt, 1996). In our model we assume a pre-specified policy for filling backordered demand (applying lateral transhipments and direct deliveries) and determine the optimal, i.e. minimal cost, allocation of stock in the system. We compare the performance of our ESM with the performance of the VARI-METRIC model in which no flexibility is applied and all demand is backordered at the central warehouse. We also use our model to evaluate an existing distribution system.

This chapter is organized as follows. In Section 5.2 we discuss some of the literature that is closely related to our model and discuss the differences. In Section 5.3 we present the Emergency Supply Model in detail and list the assumptions and notations. In Section 5.4 we analyze the model using a two-step decomposition approach and Markov analysis. Since our model is an approximate model, we validate the model performance by means of simulation in Section 5.5. Here we also present a sensitivity analysis with respect to the lead time distribution. In Section 5.6 we compare the performance of our ESM with the performance of the standard VARI-METRIC model. For this purpose we introduce a cost function that incorporates the most important cost factors. In Section 5.7 we investigate different pooling structures and their impact on the service performance of the model. In Section 5.8 we describe a case-study where we used our model to evaluate the performance of a European service parts distribution system in practice. Finally, in Section 5.9 we discuss the results and give some concluding remarks.

5.2 Related literature

A number of authors have investigated the use of pooling flexibility and/or direct shipment flexibility in multi-echelon inventory systems with one-for-one replenishment policies. In the previous chapter we already discussed the paper by Muckstadt and Thomas (1980) who extended the METRIC model to allow for direct deliveries from a higher echelon (i.e. central warehouse or production plant) in case of a stock-out at an inventory location at the lowest echelon. In this section we discuss some papers that address the option of using emergency lateral transhipments (ELT's) in multi-echelon inventory systems.

Emergency Supply Model

Lee (1987) presents a two-echelon model with one-for-one replenishment in which he allows for lateral transshipments between the local warehouses. The local warehouses are supplied from a central warehouse which in turn is supplied from the plant which is assumed to have infinite supply. The local warehouses are grouped into a number of pooling groups. Within each group the warehouses are assumed to be identical. If demand cannot be satisfied from stock on hand, ELT's are used to fill the demand from another warehouse in the same pooling group that has stock on hand. If this is not possible, the demand is backordered. Lee derives approximate expressions for the fraction of demand satisfied from stock on hand, the fraction of demand that is satisfied by ELT, and the fraction of demand that is backordered. He compares his approximations with simulation results when different sourcing rules (which local warehouse in the group will source the ELT?) are used. The results show that the differences between sourcing rules are not significant and that the approximation is accurate for high values of the fill rate (> 0.70). In his paper Lee also presents an algorithm for determining optimal stocking levels such that costs (holding, backorder, and ELT) are minimized, subject to service level constraints.

Axsäter (1990) analyzes the same system as Lee does. The local warehouses in each pooling group, however, do not have to be identical. He uses a different modelling approach by concentrating on the demand processes at the local warehouses. When stock on hand is positive, the demand faced by the local warehouse equals the normal demand plus some ELT demand from other local warehouses in the same pooling group. When stock on hand is not positive, the only demand faced by the local warehouse is the backordered demand. Steady-state probabilities are derived by assuming exponentially distributed replenishment times. The analytical results are compared with simulation results. For the case with identical warehouses in each pooling group Axsäter compares his model with Lee's model and finds better results. Also for the case with nonidentical warehouses Axsäter's model gives satisfactory results.

Sherbrooke (1992a) presents a simulation study in which he investigates the added value of using lateral transshipments in a two-echelon depot-base system for repairable items. In contrast with Lee and Axsäter, Sherbrooke allows for delayed lateral transshipments. This means that if a base has no inventory on hand and receives a replenishment order from the depot, this unit may be laterally transhipped to another base with a backorder. Sherbrooke assumes that an ELT, normal or delayed, is only issued if it will arrive sooner than a pipeline unit. Upper and lower bounds for the expected system backorders are derived. Next regression analysis on the simulation data is used to derive approximate expressions for the expected system backorders. For depot-only-repairable items Sherbrooke shows that an average backorder reduction of 30-50 % is possible (with a maximum of 72 %) when using ELT's.

Pyke (1990) presents a simulation study for a two-echelon system for repairable parts for electronic

equipment on military aircraft. His main goal is to investigate the use of priority rules for the central repair shop in conjunction with priority rules for allocating repaired items to the bases. With regard to lateral transshipments he concludes that the improvement of the performance is marginal when decreasing the lateral transshipment times. The major gain is obtained in the limit, when the lateral transshipment times go to zero.

The Emergency Supply Model presented in this chapter differs from the previous models (Muckstadt and Thomas (1980), Lee (1987), Axsäter (1990), Pyke (1990), and Sherbrooke (1992)) because we combine two flexibility options in one model: (1) lateral transshipments within the same echelon and (2) direct deliveries from a higher echelon. Muckstadt and Thomas do not allow for lateral transshipments and only consider direct deliveries from a higher echelon. Lee, Axsäter, Pyke, and Sherbrooke focus on the use of lateral transshipments and do not allow for direct deliveries. When lateral transshipments are not possible (i.e. all inventory locations at the lowest echelon are out of stock) they assume that the demand is backordered at the central depot. In our model we apply a second flexibility option for satisfying this demand by using direct deliveries.

Dada (1992) models a two-echelon system with priority shipments which is closely related to our model. Demand that can not be satisfied from stock on hand at a local warehouse is satisfied through emergency lateral transshipments or a direct delivery from the central warehouse. If this is not possible, Dada assumes that any item in transit from the central warehouse to the local warehouses can be used to satisfy this demand. Therefore Dada assumes that full information is available about items in transit and that these items can be redirected to any other destination after arrival at the original destination. Although current information systems make this sort of pipeline information more accessible, it is in most practical situations not feasible to apply this option. Examples are boats, planes, or trucks of service providers. We therefore restrict ourselves to priority shipment options from physical stock locations. Dada applies a similar modelling approach as we do. He first constructs an aggregate model for the local warehouses and next presents a disaggregation scheme to find the performance of the individual locals. However, Dada assumes in his model that all local warehouses have identical lead times and stock level one. The analysis of his model is rather complex and the approximation scheme he presents for the case of non-identical local warehouses does not always converge.

5.3 Model description

The Emergency Supply Model (ESM) is a two-echelon inventory model consisting of a central warehouse supplying a number of local warehouses. The central warehouse itself is supplied by a production plant which we assume to have infinite capacity. The inventory policy applied is

Emergency Supply Model

one-for-one replenishment for all inventory locations in the system. This policy is common practice for very expensive service parts with a low demand rate. In case of a demand at a local warehouse, we apply the following strategy for filling this demand:

- 1) Fill the demand from stock on hand. The local warehouse where the demand occurs issues a replenishment order to the central warehouse.
- 2) If no stock on hand is available, the demand is satisfied by an emergency lateral transshipment (ELT) from another, randomly chosen, local warehouse that has stock on hand. We assume that the local warehouses are situated in one geographical area such that the ELT time is much shorter than the replenishment time from the central warehouse. The local warehouse that sources the ELT issues a replenishment order to the central warehouse.
- 3) If ELT's are not possible, the demand is satisfied by a direct delivery from the central warehouse if it has stock on hand. Here we assume that a direct delivery is much faster than a normal replenishment from the central warehouse. The central warehouse issues a replenishment order to the plant.
- 4) Finally, if a direct delivery from the central warehouse is not possible, the demand is satisfied by a direct delivery from the plant which has infinite supply. Note that this last option is equivalent to modelling a lost demand or using a source outside the system.

The structure of the inventory system and the policy used for filling demand is depicted in figure 5.2.

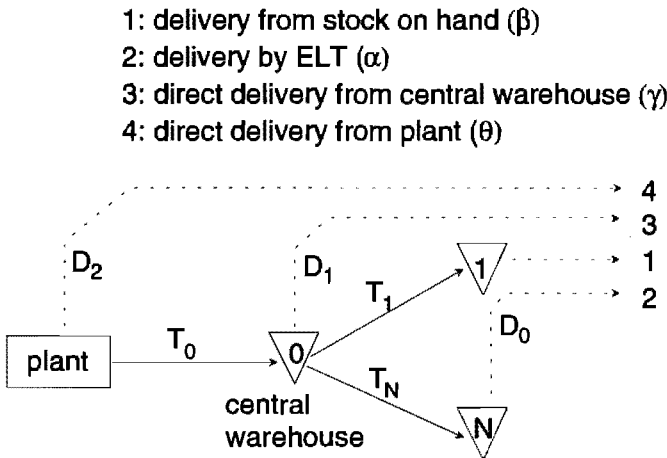


Figure 5.2: The inventory system

The number of local warehouses is denoted by N . We will use index $i, i=1\dots N$ to denote a specific local warehouse, and 0 for the central warehouse. Customers are assumed to arrive at local warehouse i according to a Poisson process with constant intensity λ_i . Furthermore, we let $\tilde{\lambda}$ denote the total arrival intensity of customers, i.e.,

$$\tilde{\lambda} = \sum_{i=1}^N \lambda_i$$

We let S_i denote the stock level at local warehouse i , and S_0 the stock level at the central warehouse. A customer arriving at local warehouse i will receive an item from stock on hand if stock is available. By β_i we denote the fraction of the demand λ_i that can be met directly from stock on hand. Since one-for-one replenishment is used, a customer served from stock on hand will trigger a replenishment order from the central warehouse to local warehouse i . If a customer arrives at local warehouse i when this warehouse is out of stock, we serve the customer by issuing an ELT from a neighboring warehouse, chosen randomly from those local warehouses with positive stock on hand. The fraction of the demand λ_i that is met by ELT is denoted by α_i . In principle, this could be seen as a redirection of the customer to another local warehouse. In particular, the central warehouse will receive an order for replenishment from the local warehouse sourcing the ELT. We assume that the customer initiating the lateral transshipment will wait for this specific item, although the local warehouse could receive items through normal replenishment while the customer is waiting. Note that the local warehouses are assumed to form one pooling group, and that the average lateral transshipment time, D_p , is identical for all transshipments between local warehouses. In the case when all local warehouses are out of stock but the central warehouse has stock on hand, an arriving customer at any local warehouse is served through direct delivery from the central warehouse. By γ we denote the fraction of the total demand $\tilde{\lambda}$ that is met in this way. It is clear from our pooling assumption that γ is also the fraction of customers arriving at local warehouse i that will be served through direct delivery from the central warehouse. As above, we assume that a customer initiating a direct delivery from the central warehouse will wait for this item to arrive, with an average direct delivery time of D_1 . If a customer arrives when there is no stock on hand, locally or at the central warehouse, the customer is served by direct delivery from the plant. The fraction of the total demand $\tilde{\lambda}$ that will be met in this way is denoted by θ . This is also the fraction of customers arriving at any local warehouse that will be satisfied through direct delivery from the plant. Again, we assume that a customer initiating a direct delivery from the plant will wait for this item to arrive, with an average direct delivery time of D_2 .

If the central warehouse is out of stock when receiving a replenishment order, the demand is backlogged. Demand, including backlogged demand, at the central warehouse is satisfied on the

Emergency Supply Model

basis of first come first served (FCFS). For local warehouse i , the time between placing a replenishment order and receiving the ordered item is called the local replenishment lead time, denoted by L_i . It consists of the transportation time T_i from the central warehouse to local warehouse i plus, due to occasional stock-outs at the central warehouse, an expected delay Δ . Since our approach is based on a Markov analysis, we assume that the local replenishment lead time at local warehouse i is exponentially distributed with mean L_i . In Section 5.5.2 of this chapter we will use simulation to test if this assumption is very restrictive for our model. We also assume that the local replenishment lead times at the local warehouses are independent of each other.

When a demand arrives at a local warehouse, this can result in an indirect demand (when the demand is satisfied from stock on hand or through ELT) or a direct demand (when the demand is satisfied through a direct delivery from the central warehouse) at the central warehouse. Consequently, the central warehouse issues a replenishment order to the production plant. The expected time the central warehouse has to wait before it receives the ordered item will be called the central replenishment lead time. As mentioned earlier, we assume that the plant has infinite supply, and therefore the central replenishment lead time equals the shipment time T_0 from the plant to the central warehouse. In our model, we assume that the shipment times are independent and exponentially distributed with mean T_0 .

A summary of the notation used in the analysis is given below:

- N : number of local warehouses
- λ_i : customer arrival rate at local warehouse i
- $\bar{\lambda}$: total customer arrival rate at all local warehouses
- S_0 : initial stock level at the central warehouse
- S_i : initial stock level at local warehouse i
- T_0 : shipment time from the plant to the central warehouse
- T_i : shipment time from the central warehouse to local warehouse i
- Δ : delay at the central warehouse
- L_i : local replenishment lead time at local warehouse i
- D_0 : ELT time between local warehouses
- D_1 : direct delivery time from the central warehouse
- D_2 : direct delivery time from the plant
- α_i : fraction of demand at local warehouse i satisfied through ELT
- β_i : fraction of demand at local warehouse i satisfied from stock on hand
- γ : fraction of total demand satisfied through direct delivery from the central warehouse
- θ : fraction of total demand satisfied through direct delivery from the plant

5.4 Model analysis

We analyze the system in two steps. First, in Section 5.4.1, we construct an aggregate model that enables us to calculate the fraction of demand that is satisfied by a direct delivery from the central warehouse (γ) and the fraction of demand that is satisfied by a direct delivery from the plant (θ). These fractions are identical for all local warehouses, even if they have different demand rates and stock levels. Second, in Section 5.4.2, we construct a model for every local warehouse separately that enables us to estimate the fraction of demand satisfied from stock on hand (β_i) and the fraction of demand satisfied through ELT (α_i). In this second step we apply a technique introduced by Axsäter (1990).

5.4.1 Aggregate model for γ and θ

In the first step of the analysis we derive expressions for γ and θ . We also derive an expression for Δ , the expected delay for local replenishment orders at the central warehouse due to occasional stock-out situations, which we will need in the second step (see Section 5.4.2). The approach, which follows closely the idea described in Dada (1992), is based on the observation that from the point of view of the central warehouse, the local warehouses behave as one aggregate warehouse with stock level \bar{S} , with

$$\bar{S} = \sum_{i=1}^N S_i$$

We construct a finite two-dimensional Markov model with states j and k :

- j = net inventory at central warehouse ($-\bar{S} \leq j \leq S_0$)
 k = net inventory at the aggregate local warehouse ($0 \leq k \leq \bar{S}$)

The associated steady state probabilities are denoted by π_{jk} . Note that the net inventory, defined as the physical stock on hand minus backorders, can never be negative for the local warehouses, since backorders are not allowed. A negative value of j corresponds to $-j$ backorders at the central warehouse. There are three events leading to a change of state: (1) a customer arrives at the aggregate local warehouse, (2) a replenishment order arrives at the aggregate local warehouse, and (3) a replenishment order arrives at the central warehouse. The rates at which these three events occur are denoted as follows:

Emergency Supply Model

- $\bar{\lambda}$ = the rate at which customers arrive at the aggregate local warehouse
- $\bar{\mu}$ = the rate at which a replenishment order arrives at the aggregate local warehouse,
- μ_0 = the rate at which a replenishment order arrives at the central warehouse.

The arrival rate $\bar{\lambda}$ of customers at the aggregate local warehouse is equal to the sum of the individual arrival rates at all local warehouses. The replenishment rate μ_0 at the central warehouse is equal to $1/T_0$, since the production plant is assumed to have infinite capacity. The replenishment rate $\bar{\mu}$ at the aggregate local warehouse deserves some further attention. If the shipment times are all exponentially distributed with the same mean, i.e. $T_i = T_j$ for all i , then $\bar{\mu}=1/T_1$, and the model is exact. If this is not the case, we use the following approximation:

$$\bar{T} = \sum_{i=1}^N \frac{\lambda_i T_i}{\bar{\lambda}}$$

$$\bar{\mu} = \frac{1}{\bar{T}}$$

That is, we assume that the rate at which items are being sent from the central warehouse to local warehouse i is equal to the demand rate λ_i . Due to lateral transshipments, this is generally not the case. Moreover, we assume that the shipment times from the central warehouse to the aggregate local warehouse is exponentially distributed with rate $\bar{\mu}$. Note that even if all individual shipment times are exponentially distributed but with different means, this will not be the case.

The state space and corresponding transition rates are depicted in figure 5.3 for the case when $S_0 = 2$ and $\bar{S} = 2$. From this picture it is clear how to form the state space for different values of S_0 and \bar{S} . We can now find the steady-state probabilities π_{jk} by solving the corresponding system of linear equations together with the normalizing constraint:

$$\sum_{j=-\bar{S}}^{S_0} \sum_{k=0}^{\min(S_0, \bar{S}-j)} \pi_{jk} = 1$$

Having found the steady-state probabilities π_{jk} and using the PASTA property, we can find γ , θ , and Δ as follows:

$$\gamma = \sum_{j=1}^{s_0} \pi_{j,0}$$

$$\theta = \sum_{j=s}^{\infty} \pi_{j,0}$$

$$\Delta = \frac{1}{(1-\theta)\bar{\lambda}} \sum_{j=s}^{-1} (-j) \sum_{k=0}^{s+j} \pi_{j,k}$$

The expression for Δ is found by applying the well-known result by Little which states that the average waiting time is equal to the expected number of backorders divided by the arrival rate.

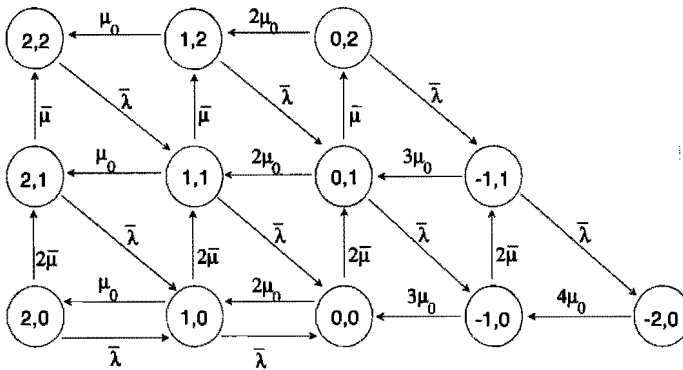


Figure 5.3: Aggregate state space model

5.4.2 Model for α_i and β_i

The objective of the second step of our heuristic is to find estimates of α_i and β_i for all local warehouses i . Again we will use Markov analysis, following closely the approach described by Axsäter (1990). The main idea is to adjust the demand rate at a local warehouse by taking into consideration that the local warehouse will sometimes be used as a source for ELT's to other local warehouses. When local warehouse i has stock on hand, it will face the regular demand with an average rate λ_i plus ELT-demand from other local warehouses with an average rate e_i (the actual expression for e_i is presented later). When the net inventory is zero the demand is zero, since customers arriving in this state are satisfied either through an ELT or a direct delivery. Let $g_i = \lambda_i + e_i$ denote the adjusted demand rate at local i . We assume that the demand process at local warehouse i is Poisson with rate g_i and that the demand processes at different local warehouses are independent.

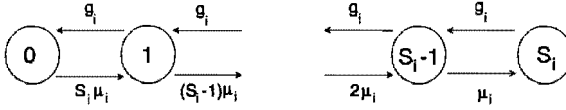


Figure 5.4: State space model for local i

As mentioned previously, we assume that the replenishment lead times for local warehouse i are independent and exponentially distributed with mean $L_i = T_i + \Delta$. The replenishment rate is then $\mu_i = 1/L_i$ and the corresponding state space is depicted in figure 5.4. The steady-state equations can be solved analytically as follows. Define:

$$p_j^i := \Pr [j \text{ items on hand at local warehouse } i], \quad 0 \leq j \leq S_i$$

Then

$$p_{S_i}^i = \left\{ \sum_{j=0}^{S_i} \frac{(g_i/\mu_i)^j}{j!} \right\}^{-1}$$

$$p_j^i = \frac{(g_i/\mu_i)^{S_i-j}}{(S_i-j)!} p_{S_i}^i, \quad (j=0 \dots S_i-1)$$

After computing these steady-state probabilities, we can readily find the estimates $\beta_i = 1 - p_0^i$ and $\alpha_i = 1 - \beta_i - \gamma - \theta$.

Still we need an expression for e_i . Let D_i denote the fraction of the total demand $\bar{\lambda}$ that is satisfied at local warehouse i . Then D_i is equal to $\lambda_i + e_i$ if warehouse i has stock on hand, and D_i is equal to zero if warehouse i is out of stock. So:

$$D_i = \beta_i(\lambda_i + e_i) + (1 - \beta_i)0$$

D_i can also be interpreted as the sum of the fraction β_i of the demand λ_i originating at local warehouse i and the fraction of the ELT-demand of all other local warehouses that is satisfied by local warehouse i . So:

$$D_i = \beta_i \lambda_i + \sum_{\substack{k=1 \\ k \neq i}}^N \omega_k^i \alpha_k \lambda_k$$

where ω_k^i represents the probability that local warehouse i acts as ELT-source for local warehouse k under the condition that ELT is possible. Note that for $N=2$ this probability is always equal to one. We now have the following expression for e_i :

$$e_i = \frac{1}{\beta_i} \sum_{\substack{k=1 \\ k \neq i}}^N \omega_k^i \alpha_k \lambda_k$$

When all local warehouses are identical with respect to demand rate, stock level, and lead time, and a randomly chosen neighbor with stock on hand is used to source an ELT, the probability ω_k^i is equal to $1/(N-1)$, and hence it follows that $e_i = \alpha_i \lambda_i / \beta_i$. Note that e_i depends on α_i and β_i . Hence, we use an iterative procedure where we alternately update the values for β_i , α_i and e_i . However, as Axsäter (1990) also notes, convergence is obtained after a few iterations. In our case, we start the iteration procedure with $\beta_i = 1 - \gamma - \theta$ and $\alpha_i = 0$, which implies that e_i is initially zero for all i .

When the local warehouses are not identical, the expressions for ω_k^i become more complicated, but the general idea remains the same, i.e. we iteratively solve for β_i , α_i , and e_i . If a randomly chosen neighbor with stock on hand is used to source an ELT, the general expression for ω_k^i is as follows ($N > 2$):

$$\begin{aligned} \omega_k^i &= \Pr\{i \text{ acts as ELT-source for } k \mid \text{ELT is possible}\} \\ &= \frac{\Pr\{\text{ELT is possible and } i \text{ acts as source for } k\}}{\Pr\{\text{ELT is possible}\}} \\ &= \frac{\beta_i \sum_{A \subseteq \{1, \dots, N\} \setminus \{i, k\}} \frac{1}{S(A)+1} \prod_{j=1}^N \{x_j \beta_j + (1-x_j)(1-\beta_j)\}}{\sum_{\substack{A \subseteq \{1, \dots, N\} \setminus \{k\} \\ A \neq \emptyset}} \prod_{\substack{j=1 \\ j \neq k}}^N \{x_j \beta_j + (1-x_j)(1-\beta_j)\}} \end{aligned}$$

where $S(A)$ represents the number of elements in set A and x_j represents the following indicator function:

$$\begin{aligned} x_j &= 1 \quad \text{if } j \in A \\ x_j &= 0 \quad \text{if } j \notin A \end{aligned}$$

For $N=2$ we have the following expressions for e_i :

$$e_1 = \frac{\alpha_2 \lambda_2}{\beta_1}, \quad e_2 = \frac{\alpha_1 \lambda_1}{\beta_2}$$

For $N=3$ the expression for e_1 looks as follows (analogously for $i=2$ and $i=3$):

$$e_1 = \frac{\alpha_2 \lambda_2 (1 - \beta_3 / 2)}{\beta_1 + \beta_3 - \beta_1 \beta_3} + \frac{\alpha_3 \lambda_3 (1 - \beta_2 / 2)}{\beta_1 + \beta_2 - \beta_1 \beta_2}$$

The formulas for e_i presented above are based on a random choice of source for an ELT. However, we would like to point out that the modelling approach could be used also in the case when each local uses a priority list for determining the source of ELT's. Just as long as the local warehouse looks to *all* other locals before using a direct delivery from the central warehouse or plant, the aggregate model presented in Section 5.4.1 is still valid. In particular, the values for γ and θ would again be exact if the shipment times are exponentially distributed with the same mean. The difference is that the formulas for e_i change, and thus the values for α_i and β_i .

5.5 Model validation

In this section, we present simulation results in order to show the accuracy of our model (Section 5.5.1). We consider a situation with three local warehouses, i.e., $N=3$ throughout this section. For each simulation run we simulated a minimum of 500.000 customer arrivals at each local warehouse. We also test the sensitivity of our model with regard to the lead time distribution (Section 5.5.2).

5.5.1 Simulation results for α_i , β_i , γ and θ

Since our modelling technique assumes exponentially distributed lead times, we first simulate the situation where the central lead time and the shipment times are drawn from exponential distributions with mean T_0 and T_i respectively.

First, we consider a set of configurations in which the three local warehouses are identical. Because we apply direct deliveries when all local warehouses are out of stock, the results from our model can not be compared directly to, e.g., the models of Lee (1987) and Axsäter (1990). However, the parameter values used in the problems are chosen similar to theirs. The central shipment time, T_0 , equals 15, and the local shipment times, T_i , are all equal to 3. The results from the simulation with exponentially distributed shipment times (exp-column) and from our emergency supply model (ESM-column) are shown in table 5.1 for γ , θ and β_i . The value for α_i can readily be obtained from the relationship $\alpha_i + \beta_i + \gamma + \theta = 1$. Note that when the shipment times are the same for all locals and independent exponentially distributed, the values for γ and θ should be exact, which is indeed the case.

| case | λ_i | S_i | S_0 | γ | | | θ | | | β_i | | |
|------|-------------|-------|-------|----------|--------------------------|--------------------------|----------|--------------------------|--------------------------|-----------|--------------------------|--------------------------|
| | | | | ESM | Simulation <i>exp</i> | Simulation <i>det</i> | ESM | Simulation <i>exp</i> | Simulation <i>det</i> | ESM | Simulation <i>exp</i> | Simulation <i>det</i> |
| 1 | 0.02 | 1 | 1 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.84 | 0.84 | 0.84 |
| 2 | | | 2 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.92 | 0.92 | 0.92 |
| 3 | 0.06 | 1 | 1 | 0.00 | 0.00 | 0.00 | 0.23 | 0.23 | 0.23 | 0.48 | 0.48 | 0.48 |
| 4 | | | 2 | 0.00 | 0.00 | 0.00 | 0.13 | 0.13 | 0.13 | 0.61 | 0.61 | 0.61 |
| 5 | | 2 | 1 | 0.00 | 0.00 | 0.00 | 0.03 | 0.03 | 0.03 | 0.83 | 0.81 | 0.81 |
| 6 | | | 2 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.91 | 0.89 | 0.89 |
| 7 | 0.10 | 1 | 2 | 0.00 | 0.00 | 0.00 | 0.32 | 0.32 | 0.32 | 0.40 | 0.40 | 0.40 |
| 8 | | | 3 | 0.00 | 0.00 | 0.00 | 0.23 | 0.23 | 0.22 | 0.49 | 0.49 | 0.49 |
| 9 | | | 6 | 0.02 | 0.02 | 0.02 | 0.06 | 0.06 | 0.06 | 0.67 | 0.67 | 0.67 |
| 10 | | | 10 | 0.05 | 0.05 | 0.05 | 0.00 | 0.00 | 0.00 | 0.71 | 0.71 | 0.71 |
| 11 | | 2 | 2 | 0.00 | 0.00 | 0.00 | 0.09 | 0.09 | 0.09 | 0.71 | 0.69 | 0.69 |
| 12 | | | 3 | 0.00 | 0.00 | 0.00 | 0.05 | 0.05 | 0.05 | 0.80 | 0.78 | 0.78 |
| 13 | | 3 | 2 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.90 | 0.88 | 0.88 |
| 14 | | | 3 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.95 | 0.93 | 0.93 |

Table 5.1: Simulation results for identical local warehouses, $N = 3$, $T_0 = 15$, and $T_i = 3$. (*exp*=exponentially distributed lead times; *det*=deterministic lead times)

Second, we also consider a set of configurations in which the local warehouses are not identical, but have different demand rates λ_i , or different shipment times T_i . The results are shown in table 5.2 and table 5.3. For the case with identical locals (table 5.1) our model gives excellent results. When the local stock levels are equal to one our model even yields exact results. For the case with non-identical locals (tables 5.2 and 5.3) the model still performs very well.

5.5.2 Sensitivity to lead time distribution

Our modelling assumptions include exponentially distributed shipment times for normal replenishments. However, in practice these shipment times are often close to deterministic since they consist of transportation times between inventory locations. Therefore, we also simulated the system with deterministic lead times. The results are also presented in tables 5.1, 5.2 and 5.3 (*det*-column). An analysis of these results shows that the service performance of the system is almost identical for exponential and deterministic shipment times. Apparently the lead time distribution does not affect the service performance. In fact, the key METRIC assumption is that Palm's theorem for infinite server queues applies to the replenishment process of the local warehouses as well. Palm's theorem states that the distribution of parts in resupply is only dependent on the replenishment time distribution through its mean. The results indicate that our model is to a large extent insensitive to the choice of the lead time distribution.

| λ_i | T_i | S_i | S_0 | γ | | | θ | | | β_i | | |
|-------------|-------|-------|-------|----------|--------------------------|--------------------------|----------|--------------------------|--------------------------|-----------|--------------------------|--------------------------|
| | | | | ESM | Simulation <i>exp</i> | Simulation <i>det</i> | ESM | Simulation <i>exp</i> | Simulation <i>det</i> | ESM | Simulation <i>exp</i> | Simulation <i>det</i> |
| 0.01 | 3 | 1 | 1 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.89 | 0.88 | 0.88 |
| 0.02 | 3 | | | | | | | | | 0.84 | 0.84 | 0.84 |
| 0.03 | 3 | | | | | | | | | 0.80 | 0.81 | 0.81 |
| 0.01 | 3 | 1 | 2 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.95 | 0.94 | 0.94 |
| 0.02 | 3 | | | | | | | | | 0.91 | 0.91 | 0.91 |
| 0.03 | 3 | | | | | | | | | 0.88 | 0.89 | 0.89 |
| 0.02 | 1 | 1 | 1 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.88 | 0.88 | 0.87 |
| 0.02 | 3 | | | | | | | | | 0.84 | 0.84 | 0.84 |
| 0.02 | 5 | | | | | | | | | 0.82 | 0.82 | 0.82 |
| 0.02 | 1 | 1 | 2 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.95 | 0.95 | 0.95 |
| 0.02 | 3 | | | | | | | | | 0.92 | 0.92 | 0.92 |
| 0.02 | 5 | | | | | | | | | 0.88 | 0.88 | 0.88 |
| 0.02 | 3 | 1 | 2 | 0.00 | 0.00 | 0.00 | 0.13 | 0.13 | 0.13 | 0.69 | 0.67 | 0.66 |
| 0.06 | 3 | | | | | | | | | 0.61 | 0.61 | 0.61 |
| 0.10 | 3 | | | | | | | | | 0.54 | 0.56 | 0.56 |
| 0.02 | 3 | 2 | 1 | 0.00 | 0.00 | 0.00 | 0.03 | 0.03 | 0.03 | 0.92 | 0.88 | 0.88 |
| 0.06 | 3 | | | | | | | | | 0.81 | 0.80 | 0.80 |
| 0.10 | 3 | | | | | | | | | 0.72 | 0.73 | 0.73 |
| 0.02 | 3 | 2 | 2 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.97 | 0.94 | 0.94 |
| 0.06 | 3 | | | | | | | | | 0.90 | 0.89 | 0.89 |
| 0.10 | 3 | | | | | | | | | 0.83 | 0.83 | 0.83 |
| 0.06 | 1 | 1 | 2 | 0.00 | 0.00 | 0.00 | 0.13 | 0.13 | 0.13 | 0.68 | 0.68 | 0.67 |
| 0.06 | 3 | | | | | | | | | 0.62 | 0.61 | 0.61 |
| 0.06 | 5 | | | | | | | | | 0.57 | 0.57 | 0.57 |
| 0.06 | 1 | 2 | 1 | 0.00 | 0.00 | 0.00 | 0.03 | 0.03 | 0.03 | 0.86 | 0.84 | 0.84 |
| 0.06 | 3 | | | | | | | | | 0.83 | 0.81 | 0.81 |
| 0.06 | 5 | | | | | | | | | 0.80 | 0.78 | 0.78 |
| 0.06 | 1 | 2 | 2 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.94 | 0.92 | 0.92 |
| 0.06 | 3 | | | | | | | | | 0.91 | 0.89 | 0.89 |
| 0.06 | 5 | | | | | | | | | 0.88 | 0.87 | 0.87 |

Table 5.2: Simulation results for non-identical local warehouses, $N = 3$ and $T_0 = 15$.
 (*exp*=exponentially distributed lead times; *det*=deterministic lead times)

| λ_1 | T_1 | S_1 | S_0 | γ | | | θ | | | β_1 | | |
|-------------|-------|-------|-------|----------|--------------------------|--------------------------|----------|--------------------------|--------------------------|-----------|--------------------------|--------------------------|
| | | | | ESM | Simulation <i>exp</i> | Simulation <i>det</i> | ESM | Simulation <i>exp</i> | Simulation <i>det</i> | ESM | Simulation <i>exp</i> | Simulation <i>det</i> |
| 0.05 | 3 | 1 | 6 | 0.02 | 0.02 | 0.02 | 0.06 | 0.06 | 0.06 | 0.73 | 0.73 | 0.73 |
| 0.10 | 3 | | | | | | | | | 0.67 | 0.67 | 0.67 |
| 0.15 | 3 | | | | | | | | | 0.61 | 0.62 | 0.62 |
| 0.05 | 3 | 1 | 10 | 0.05 | 0.05 | 0.05 | 0.00 | 0.00 | 0.00 | 0.77 | 0.77 | 0.77 |
| 0.10 | 3 | | | | | | | | | 0.71 | 0.71 | 0.71 |
| 0.15 | 3 | | | | | | | | | 0.66 | 0.66 | 0.66 |
| 0.05 | 3 | 2 | 2 | 0.00 | 0.00 | 0.00 | 0.09 | 0.09 | 0.09 | 0.79 | 0.74 | 0.74 |
| 0.10 | 3 | | | | | | | | | 0.70 | 0.68 | 0.68 |
| 0.15 | 3 | | | | | | | | | 0.63 | 0.63 | 0.63 |
| 0.05 | 3 | 2 | 3 | 0.00 | 0.00 | 0.00 | 0.05 | 0.05 | 0.05 | 0.88 | 0.83 | 0.83 |
| 0.10 | 3 | | | | | | | | | 0.80 | 0.77 | 0.77 |
| 0.15 | 3 | | | | | | | | | 0.72 | 0.72 | 0.72 |
| 0.10 | 1 | 1 | 6 | 0.02 | 0.01 | 0.01 | 0.06 | 0.06 | 0.05 | 0.82 | 0.81 | 0.81 |
| 0.10 | 3 | | | | | | | | | 0.69 | 0.68 | 0.68 |
| 0.10 | 5 | | | | | | | | | 0.59 | 0.59 | 0.59 |
| 0.10 | 1 | 1 | 10 | 0.05 | 0.03 | 0.03 | 0.00 | 0.00 | 0.00 | 0.88 | 0.87 | 0.87 |
| 0.10 | 3 | | | | | | | | | 0.73 | 0.72 | 0.72 |
| 0.10 | 5 | | | | | | | | | 0.63 | 0.62 | 0.62 |
| 0.10 | 1 | 2 | 2 | 0.00 | 0.00 | 0.00 | 0.09 | 0.09 | 0.09 | 0.76 | 0.73 | 0.73 |
| 0.10 | 3 | | | | | | | | | 0.71 | 0.69 | 0.69 |
| 0.10 | 5 | | | | | | | | | 0.67 | 0.65 | 0.65 |
| 0.10 | 1 | 2 | 3 | 0.00 | 0.00 | 0.00 | 0.05 | 0.05 | 0.05 | 0.86 | 0.82 | 0.82 |
| 0.10 | 3 | | | | | | | | | 0.80 | 0.78 | 0.78 |
| 0.10 | 5 | | | | | | | | | 0.76 | 0.74 | 0.74 |

Table 5.3: Simulation results for non-identical local warehouses, $N = 3$ and $T_0 = 15$.
(*exp*=exponentially distributed lead times; *det*=deterministic lead times)

5.6 Economic evaluation

In order to evaluate the economic performance of the model we need a cost structure that takes into account the different operational cost factors. These include: inventory holding costs, normal replenishment costs (to the central warehouse and to the local warehouses), emergency shipment costs (ELT, direct delivery from the central warehouse and direct delivery from the production plant) and penalty costs for customers who have to wait. We need the following cost parameters to make an economic evaluation:

- c : unit price of one item
- h_0 : inventory holding cost at the central warehouse per item per time unit expressed as a fraction of the unit price
- h_i : inventory holding cost at local warehouse i per item per time unit expressed as a fraction of the unit price
- r_0 : normal replenishment cost per item at the central warehouse
- r_i : normal replenishment cost per item at local warehouse i
- s_0 : emergency replenishment cost for using an ELT
- s_1 : emergency replenishment cost for using a direct delivery from the central warehouse
- s_2 : emergency replenishment cost for using a direct delivery from the plant
- z_i : penalty cost per time unit for a waiting customer at local warehouse i

We assume that the cost for using a particular kind of emergency shipment is equal for all locals. However, it is possible to differentiate these costs for the different locals. The expected waiting time w_i for an arbitrary customer at local i , can be expressed as follows (using the PASTA property):

$$w_i = \alpha_i D_0 + \gamma D_1 + \theta D_2$$

Given the steady-state probabilities π_{jk} and p_j^i we calculated in Section 5.4, we can now formulate the total cost TC per time unit as follows:

$$\begin{aligned}
 TC = & c h_0 \sum_{j=1}^{S_0} \sum_{k=0}^S j \pi_{jk} + c \sum_{i=1}^N h_i \sum_{j=1}^{S_i} j p_j^i + e_0 \sum_{i=1}^N \alpha_i \lambda_i + e_1 \gamma \bar{\lambda} \\
 & + e_2 \theta \bar{\lambda} + r_0 (1 - \theta) \bar{\lambda} + \sum_{i=1}^N r_i (1 - \theta - \gamma) \lambda_i + \sum_{i=1}^N \lambda_i w_i z_i
 \end{aligned}$$

Our aim is to find the optimal stock levels for the central warehouse and the local warehouses (S_0^* and S_i^*) that minimize the total cost function presented above. The minimum cost, associated with these optimal stock levels, is denoted by TC^* . In our numerical experiment we are primarily

interested in the influence of certain parameters on the minimum total cost and optimal stock levels, namely:

- 1) The daily demand rates at the local warehouses; $\lambda_i \in \{ 0.02, 0.06, 0.10 \}$
- 2) The inventory holding cost fractions; $h_i \in \{ 0.10/365, 0.30/365 \}$
- 3) The emergency replenishment costs; $e_{0,1,2} \in \{ (10, 30, 100), (30, 90, 300) \}$
- 4) The penalty cost; $z_i \in \{ 100, 1000 \}$

The inventory holding costs are identical for all stocking locations and are 10 % and 30 % respectively of the unit price c per year. The remaining system parameters are fixed as follows:

$$\begin{aligned}
 N &= 3; \\
 c &= 10.000; \\
 r_0 &= r_i = 10; \\
 T_0 &= 15 \text{ days}; T_i = 3 \text{ days}; \\
 D_0 &= 0.5 \text{ day}; D_i = 1 \text{ day}; D_2 = 2 \text{ days}.
 \end{aligned}$$

The minimum cost (TC^*) and the optimal stock levels (S_0^* and S_i^*) are presented in table 5.4 for 24 different cases. Table 5.4 also shows the minimum total cost (TC^*) and the optimal stock levels (S_0^* , S_i^*) if no emergency supply flexibility exists in the system. The cost structure is the same as before except for emergency transportation costs that will not occur in this case. We calculated the relevant costs using the VARI-METRIC technique as described by Graves (1985). We see that in all the 24 cases we analyzed, the policy of using emergency supply flexibility as described in this paper results in a lower total cost. The right-most column of table 5.4 shows the relative decrease in costs when using emergency supply flexibility. A maximum cost reduction of 43.9 % and a minimum of 13.2 % is obtained. The results also show that in most cases the stock levels are lower when using emergency supply flexibility. Especially the central stock level shows a significant decrease.

| λ_i | h_i | $e_{0,1,2}$ | z_i | TC* | S_0^* | S_i^* | TC ^v | S_0^v | S_i^v | % |
|-------------|-------|-------------|-------|-------|---------|---------|-----------------|---------|---------|------|
| 0.02 | 0.10 | (10,30,100) | 100 | 8.84 | 0 | 1 | 13.55 | 2 | 1 | 34.8 |
| | | | 1000 | 15.04 | 2 | 1 | 21.01 | 2 | 2 | 28.4 |
| | 0.30 | (30,90,300) | 100 | 10.03 | 0 | 1 | 13.55 | 2 | 1 | 26.0 |
| | | | 1000 | 15.19 | 2 | 1 | 21.01 | 2 | 2 | 27.7 |
| | | (10,30,100) | 100 | 17.27 | 1 | 0 | 30.76 | 1 | 1 | 43.9 |
| | | | 1000 | 32.03 | 1 | 1 | 48.92 | 3 | 1 | 34.5 |
| 0.06 | 0.10 | (30,90,300) | 100 | 20.99 | 0 | 1 | 30.76 | 1 | 1 | 31.8 |
| | | | 1000 | 32.43 | 1 | 1 | 48.92 | 3 | 1 | 33.7 |
| | 0.30 | (10,30,100) | 100 | 17.22 | 1 | 2 | 23.28 | 3 | 2 | 26.0 |
| | | | 1000 | 24.86 | 3 | 2 | 31.34 | 5 | 2 | 20.7 |
| | | (30,90,300) | 100 | 18.81 | 1 | 2 | 23.28 | 3 | 2 | 19.2 |
| | | | 1000 | 25.08 | 4 | 2 | 31.34 | 5 | 2 | 20.0 |
| 0.10 | 0.10 | (10,30,100) | 100 | 30.97 | 1 | 1 | 46.94 | 4 | 1 | 34.0 |
| | | | 1000 | 54.72 | 2 | 2 | 73.90 | 5 | 2 | 26.0 |
| | 0.30 | (30,90,300) | 100 | 36.8 | 2 | 1 | 46.94 | 4 | 1 | 21.5 |
| | | | 1000 | 55.44 | 2 | 2 | 73.90 | 5 | 2 | 25.0 |
| | | (10,30,100) | 100 | 23.11 | 4 | 2 | 29.01 | 6 | 2 | 20.3 |
| | | | 1000 | 32.03 | 4 | 3 | 38.34 | 7 | 3 | 16.5 |
| 0.10 | 0.10 | (30,90,300) | 100 | 25.18 | 5 | 2 | 29.01 | 6 | 2 | 13.2 |
| | | | 1000 | 32.21 | 5 | 3 | 38.34 | 7 | 3 | 16.0 |
| | 0.30 | (10,30,100) | 100 | 42.54 | 2 | 2 | 62.01 | 5 | 2 | 31.4 |
| | | | 1000 | 71.28 | 3 | 3 | 92.86 | 6 | 3 | 23.2 |
| | | (30,90,300) | 100 | 48.76 | 3 | 2 | 62.01 | 5 | 2 | 21.4 |
| | | | 1000 | 71.91 | 3 | 3 | 92.86 | 6 | 3 | 22.6 |

Table 5.4: Cost evaluation

5.7 Pooling structures

A key assumption in the modeling of the Emergency Supply Model is the assumption of *complete pooling*: if a local warehouse can not satisfy a demand directly from stock on hand, *all* other local warehouses are checked for the possibility of an emergency lateral transshipment. A direct delivery from the central warehouse or the factory is only made when all local warehouses are out of stock. We need this assumption in our modeling approach because we consider the local warehouses as one aggregate inventory location in the first step of our modeling technique (see Section 5.4.1). In practice, however, different pooling structures can exist. Due to geographical conditions or customs tariffs it is very well possible that only local warehouses in the 'neighborhood' are used for lateral shipments. In this section we consider two alternative pooling structures, *fixed pooling* and *variable pooling*, and investigate their impact on the service performance.

Fixed pooling

The pooling structure defined as 'fixed pooling' assumes that all local warehouses are divided into a number of fixed groups. If a local warehouse is out of stock, it can only appeal to local warehouses in the same group for an emergency lateral transshipment. In practice, fixed pooling is often applied when the market is divided into geographical areas. For example, the Scandinavian countries are usually considered as one geographical area within the European market.

Variable pooling

The pooling structure defined as 'variable pooling' assumes that all local warehouses are located in a one-dimensional space, such that each local warehouse has exactly two neighboring local warehouses. In case of a stock-out, a local warehouse can only appeal to these two neighbors for an emergency lateral transshipment. This situation can be seen in practice when the local warehouses are situated on a circle around a central warehouse.

In the following numerical experiment we compare the service performance of the Emergency Supply Model under three different pooling structures: complete pooling, fixed pooling, and variable pooling. We consider the situation with one central warehouse and nine local warehouses ($N=9$). The local warehouses are identical with respect to demand rate, stock level, and lead times. The local demand rates and the stock levels are varied in the experiment. The following parameters are fixed throughout the experiment:

$$\begin{array}{ll} T_i = 3 \text{ days} & D_o = 0.5 \text{ days} \\ T_o = 15 \text{ days} & D_1 = 1 \text{ day} \\ & D_2 = 2 \text{ days} \end{array}$$

When applying complete pooling, all local warehouses are member of one big pooling group. When applying fixed pooling, all local warehouses are divided into three non-overlapping groups with each three local warehouses. Finally, when applying variable pooling, all local warehouses are divided into nine overlapping groups with each three local warehouses. So:

| <i>Pooling structure</i> | <i>Pooling groups</i> |
|--------------------------|---|
| complete | {1,2,3,4,5,6,7,8,9} |
| fixed | {1,2,3},{4,5,6},{7,8,9} |
| variable | {1,2,3},{2,3,4},{3,4,5},{4,5,6},{5,6,7}, {6,7,8},{7,8,9},{8,9,1},{9,1,2} |

Emergency Supply Model

The simulation results are displayed in table 5.5 ($\lambda=0.02$), table 5.6 ($\lambda=0.06$), and table 5.7 ($\lambda=0.10$). The fractions α , β , γ , and θ are displayed and the average response time (RT) for backordered customers.

| λ_i | S_i | S_0 | pooling | α | β | γ | θ | RT | |
|-------------|----------|-------|----------|----------|---------|----------|----------|-------|------|
| 0.02 | 1 | 1 | complete | 0.251 | 0.748 | 0.000 | 0.001 | 0.51 | |
| | | | fixed | 0.200 | 0.762 | 0.000 | 0.038 | 0.74 | |
| | | | variable | 0.210 | 0.760 | 0.000 | 0.030 | 0.69 | |
| | | 2 | complete | 0.167 | 0.833 | 0.000 | 0.000 | 0.50 | |
| | | | fixed | 0.143 | 0.839 | 0.000 | 0.019 | 0.67 | |
| | | | variable | 0.148 | 0.838 | 0.000 | 0.014 | 0.63 | |
| | | 3 | complete | 0.113 | 0.887 | 0.000 | 0.000 | 0.50 | |
| | | | fixed | 0.101 | 0.890 | 0.000 | 0.008 | 0.61 | |
| | | | variable | 0.104 | 0.890 | 0.000 | 0.006 | 0.58 | |
| | 2 | 1 | complete | 0.028 | 0.972 | 0.000 | 0.000 | 0.50 | |
| | | | fixed | 0.028 | 0.972 | 0.000 | 0.000 | 0.51 | |
| | | | variable | 0.028 | 0.972 | 0.000 | 0.000 | 0.50 | |
| | | 2 | complete | 0.015 | 0.985 | 0.000 | 0.000 | 0.50 | |
| | | | fixed | 0.015 | 0.985 | 0.000 | 0.000 | 0.51 | |
| | | | variable | 0.016 | 0.984 | 0.000 | 0.000 | 0.50 | |
| | | 3 | complete | 0.008 | 0.992 | 0.000 | 0.000 | 0.50 | |
| | | | fixed | 0.008 | 0.992 | 0.000 | 0.000 | 0.50 | |
| | | | variable | 0.008 | 0.992 | 0.000 | 0.000 | 0.50 | |
| | | 3 | 1 | complete | 0.003 | 0.997 | 0.000 | 0.000 | 0.50 |
| | | | | fixed | 0.003 | 0.997 | 0.000 | 0.000 | 0.50 |
| | | | | variable | 0.003 | 0.997 | 0.000 | 0.000 | 0.50 |
| | | | 2 | complete | 0.001 | 0.999 | 0.000 | 0.000 | 0.50 |
| | | | | fixed | 0.001 | 0.999 | 0.000 | 0.000 | 0.50 |
| | | | | variable | 0.001 | 0.999 | 0.000 | 0.000 | 0.50 |
| 3 | complete | | 0.000 | 1.000 | 0.000 | 0.000 | 0.50 | | |
| | fixed | | 0.000 | 1.000 | 0.000 | 0.000 | 0.50 | | |
| | variable | | 0.000 | 1.000 | 0.000 | 0.000 | 0.50 | | |

Table 5.5: Service performance under different pooling concepts for $\lambda=0.02$

| λ_i | S_i | S_0 | pooling | α | β | γ | θ | RT | | |
|-------------|----------|----------|----------|----------|----------|----------|----------|-------|-------|------|
| 0.06 | 1 | 1 | complete | 0.550 | 0.253 | 0.000 | 0.197 | 0.90 | | |
| | | | fixed | 0.309 | 0.377 | 0.000 | 0.313 | 1.26 | | |
| | | | variable | 0.340 | 0.362 | 0.000 | 0.299 | 1.20 | | |
| | | 2 | complete | fixed | 0.545 | 0.308 | 0.000 | 0.147 | 0.82 | |
| | | | | fixed | 0.307 | 0.432 | 0.000 | 0.261 | 1.19 | |
| | | | | variable | 0.337 | 0.417 | 0.000 | 0.247 | 1.13 | |
| | | 4 | complete | fixed | 0.484 | 0.444 | 0.000 | 0.072 | 0.69 | |
| | | | | fixed | 0.285 | 0.547 | 0.000 | 0.168 | 1.06 | |
| | | | | variable | 0.312 | 0.534 | 0.000 | 0.154 | 1.00 | |
| | 6 | complete | fixed | 0.382 | 0.589 | 0.000 | 0.029 | 0.61 | | |
| | | | fixed | 0.250 | 0.654 | 0.001 | 0.095 | 0.91 | | |
| | | | variable | 0.265 | 0.648 | 0.001 | 0.086 | 0.87 | | |
| | 2 | 1 | complete | fixed | 0.271 | 0.726 | 0.000 | 0.003 | 0.52 | |
| | | | | fixed | 0.208 | 0.746 | 0.000 | 0.047 | 0.78 | |
| | | | | variable | 0.223 | 0.741 | 0.000 | 0.036 | 0.71 | |
| | | 2 | complete | fixed | 0.225 | 0.773 | 0.000 | 0.001 | 0.51 | |
| | | | | fixed | 0.181 | 0.786 | 0.000 | 0.033 | 0.73 | |
| | | | | variable | 0.190 | 0.785 | 0.000 | 0.025 | 0.67 | |
| | | 4 | complete | fixed | 0.142 | 0.858 | 0.000 | 0.000 | 0.50 | |
| | | | | fixed | 0.123 | 0.863 | 0.000 | 0.014 | 0.66 | |
| | | | | variable | 0.129 | 0.860 | 0.000 | 0.010 | 0.61 | |
| | | 6 | complete | fixed | 0.081 | 0.919 | 0.000 | 0.000 | 0.50 | |
| | | | | fixed | 0.072 | 0.922 | 0.000 | 0.006 | 0.61 | |
| | | | | variable | 0.077 | 0.919 | 0.000 | 0.004 | 0.57 | |
| | | 3 | 1 | complete | fixed | 0.068 | 0.932 | 0.000 | 0.000 | 0.50 |
| | | | | | fixed | 0.065 | 0.932 | 0.000 | 0.003 | 0.56 |
| | | | | | variable | 0.066 | 0.932 | 0.000 | 0.001 | 0.53 |
| | | | 2 | complete | fixed | 0.052 | 0.948 | 0.000 | 0.000 | 0.50 |
| | | | | | fixed | 0.050 | 0.948 | 0.000 | 0.002 | 0.55 |
| | | | | | variable | 0.051 | 0.948 | 0.000 | 0.001 | 0.52 |
| 4 | complete | | fixed | 0.027 | 0.973 | 0.000 | 0.000 | 0.50 | | |
| | | | fixed | 0.027 | 0.972 | 0.000 | 0.000 | 0.52 | | |
| | | | variable | 0.028 | 0.972 | 0.000 | 0.000 | 0.51 | | |
| 6 | complete | | fixed | 0.014 | 0.986 | 0.000 | 0.000 | 0.50 | | |
| | | | fixed | 0.014 | 0.986 | 0.000 | 0.000 | 0.50 | | |
| | | | variable | 0.014 | 0.986 | 0.000 | 0.000 | 0.51 | | |

Table 5.6: Service performance under different pooling concepts for $\lambda=0.06$

| λ_1 | S_1 | S_0 | pooling | α | β | γ | θ | RT |
|-------------|-------|----------|----------|----------|---------|----------|----------|------|
| 0.10 | 1 | 1 | complete | 0.443 | 0.116 | 0.000 | 0.441 | 1.25 |
| | | | fixed | 0.259 | 0.234 | 0.000 | 0.507 | 1.49 |
| | | | variable | 0.278 | 0.221 | 0.000 | 0.501 | 1.46 |
| | | 4 | complete | 0.512 | 0.191 | 0.000 | 0.297 | 1.05 |
| | | | fixed | 0.296 | 0.329 | 0.000 | 0.375 | 1.34 |
| | | | variable | 0.320 | 0.314 | 0.000 | 0.367 | 1.30 |
| | | 8 | complete | 0.513 | 0.344 | 0.000 | 0.142 | 0.83 |
| | | | fixed | 0.296 | 0.482 | 0.002 | 0.221 | 1.14 |
| | | | variable | 0.323 | 0.466 | 0.001 | 0.210 | 1.09 |
| | 12 | complete | 0.416 | 0.537 | 0.000 | 0.047 | 0.65 | |
| | | fixed | 0.266 | 0.620 | 0.011 | 0.103 | 0.92 | |
| | | variable | 0.290 | 0.606 | 0.008 | 0.096 | 0.88 | |
| | 2 | 1 | complete | 0.524 | 0.390 | 0.000 | 0.087 | 0.71 |
| | | | fixed | 0.296 | 0.515 | 0.000 | 0.189 | 1.08 |
| | | | variable | 0.333 | 0.496 | 0.000 | 0.171 | 1.01 |
| | | 4 | complete | 0.428 | 0.539 | 0.000 | 0.033 | 0.61 |
| | | | fixed | 0.263 | 0.627 | 0.000 | 0.111 | 0.94 |
| | | | variable | 0.292 | 0.611 | 0.000 | 0.097 | 0.87 |
| | | 8 | complete | 0.251 | 0.744 | 0.000 | 0.006 | 0.53 |
| | | | fixed | 0.180 | 0.777 | 0.000 | 0.043 | 0.79 |
| | | | variable | 0.194 | 0.771 | 0.000 | 0.035 | 0.73 |
| | 12 | complete | 0.114 | 0.886 | 0.000 | 0.001 | 0.51 | |
| | | fixed | 0.096 | 0.893 | 0.000 | 0.011 | 0.65 | |
| | | variable | 0.100 | 0.891 | 0.000 | 0.008 | 0.61 | |
| | 3 | 1 | complete | 0.244 | 0.754 | 0.000 | 0.002 | 0.51 |
| | | | fixed | 0.187 | 0.775 | 0.000 | 0.038 | 0.75 |
| | | | variable | 0.202 | 0.770 | 0.000 | 0.028 | 0.68 |
| | | 4 | complete | 0.153 | 0.847 | 0.000 | 0.000 | 0.50 |
| | | | fixed | 0.132 | 0.852 | 0.000 | 0.015 | 0.66 |
| | | | variable | 0.139 | 0.850 | 0.000 | 0.011 | 0.61 |
| | | 8 | complete | 0.066 | 0.934 | 0.000 | 0.000 | 0.50 |
| | | | fixed | 0.062 | 0.935 | 0.000 | 0.003 | 0.58 |
| | | | variable | 0.063 | 0.935 | 0.000 | 0.002 | 0.55 |
| | | 12 | complete | 0.022 | 0.978 | 0.000 | 0.000 | 0.50 |
| | | | fixed | 0.022 | 0.978 | 0.000 | 0.000 | 0.53 |
| | | | variable | 0.022 | 0.978 | 0.000 | 0.000 | 0.53 |

Table 5.7: Service performance under different pooling concepts for $\lambda=0.10$

The fraction of demand at a local warehouse that is satisfied through lateral transshipments, denoted by α , is maximal when applying the complete pooling concept and minimal when applying the fixed pooling concept. This is logical since under the complete pooling concept all local warehouses are candidates for sourcing an emergency lateral transshipment in case of a stock-out at an arbitrary local warehouse. When applying the fixed pooling concept only two other local warehouses (i.e. the warehouses within the same pooling group) are candidates for a lateral transshipment. Consequently, the probability that a demand is satisfied through a lateral transshipment is smaller. Application of the variable pooling concept results in an intermediate value of α : lower than under complete pooling but higher than under fixed pooling. Under the variable pooling concept, two other local warehouses (i.e. the direct neighboring warehouses) are candidates for sourcing a lateral transshipment, analogous to the fixed pooling concept. The fact that the α value is slightly higher under variable pooling than under fixed pooling is caused by the flexible use of pooling groups. This can be explained as follows. Suppose a lateral transshipment is applied in a fixed pooling group. The next stock-out event at a local warehouse in that pooling group is satisfied through a lateral transshipment with a very low probability. In case of a variable pooling group, other local warehouses are likely to be involved in subsequent stock-out events, and therefore there is a higher probability of demand satisfaction through lateral transshipments. However, it must be noted that the differences between the variable pooling concept and the fixed pooling concept are negligible for all situations. We also see that the difference between the complete pooling concept and the variable pooling concept is negligible when the fill rates β are high, i.e. higher than 70 %.

With respect to the fill rate performance, expressed as the fraction β of demand satisfied from stock on hand, we can make similar observations. The complete pooling concept results in a minimal fill rate performance, since every local warehouse is a possible source for lateral transshipments to any other local warehouse. Under the fixed pooling concept only two other local warehouses can make a request for a lateral transshipment, and consequently the fill rate performance is maximal. The variable pooling concept results in an intermediate value for β . Again we see that the differences between the fixed and variable pooling concept is negligible. The fill rate performance under the complete pooling concept is only significantly lower for low fill rates, i.e. lower than 70 %.

The value of γ (fraction of demand satisfied through a direct delivery from the central warehouse) is equal to zero in almost every case. The values of the stock levels are such that all central stock is apparently used for replenishing the local warehouses. There is only one case (table 5.7, $S_r=1$, $S_o=12$, fixed and variable pooling) in which one percent of the local demand is satisfied through direct deliveries from the central warehouse.

The value of θ (fraction of demand satisfied through a direct delivery from the factory) is minimal when applying the complete pooling concept (i.e. many lateral transshipments take place) and

Emergency Supply Model

maximal when applying the fixed pooling concept (i.e. few lateral transshipments take place). Again, the differences between fixed pooling and variable pooling are negligible and the θ -value is only significantly lower under complete pooling for low fill rates.

The average response time RT for backordered demand is calculated from the service fractions and the corresponding emergency lead times as follows: $RT = \alpha D_0 + \gamma D_1 + \theta D_2$. The results show that complete pooling gives the lowest overall response time (RT=0.60), followed by variable pooling (RT=0.73) and fixed pooling (RT=0.76).

5.8 A practical example

We now present a case study in which we applied our Emergency Supply Model to evaluate the performance of a European Service Part Supply System in practice. In Section 5.8.1 we describe the distribution network structure that is used to supply service parts from a central warehouse located in the Netherlands to countries all over Europe. In Section 5.8.2 we give an overview of the (normalized) values of the relevant model parameters. Next we selected eight pilot parts and calculated the model performance of these specific parts. The performance results of these pilot parts can be found in Section 5.8.3. We then used our model to calculate the optimal (i.e. cost minimizing) stock levels for these pilot parts. The results of this optimization step can be found in Section 5.8.4. Finally, in Section 5.8.5 we discuss some economic trade-offs that can be made with this model.

5.8.1 Network structure

The Service Part Supply System of the company in this case study consists of a central warehouse located in the Netherlands, and 19 national warehouses located in Europe, the Middle East, and Africa. In this case study we concentrated on the European service parts business and selected the following seven countries for our analysis: Finland, France, Germany, Norway, Spain, Sweden, and the United Kingdom (UK). The central warehouse is supplied by approximately 30 production plants in Europe and the USA and more than 50 external vendors all over the world. The number of different parts in the assortment is approximately 90.000. The central warehouse receives annually around 800.000 replenishment orders (either normal or emergency) from the national warehouses. In our two-echelon model we assume that the national warehouses are located at the lowest level in the supply system. However, in reality a more refined network of stocking locations exists. The national warehouses supply regional warehouses (or branch offices) from where service engineers pick up service parts and visit customers. In Section 5.8.3 we return to this issue of using a two-echelon model for analyzing a three- (or four-) echelon Service Part Supply System.

The normal replenishment procedure is a one-for-one ordering policy. In case of a demand that is met from stock on hand, the regional warehouse issues a normal replenishment order to the national warehouse in the same country, which in its turn issues a normal replenishment order to the central European warehouse. Finally, the central warehouse issues a normal replenishment order to the supplier, either a production plant or an external vendor. For fast moving service parts the central warehouse consolidates the normal replenishment orders from the different national warehouses. This consolidation of replenishment orders (for fast moving service parts) at the central warehouse is supported by comparing the number of normal replenishment orders from the central warehouse to the national warehouses (approximately 800.000 per year) with the number of normal replenishment orders from the suppliers to the central warehouse (approximately 150.000 per year).

In case of a stock-out situation at a regional warehouse several emergency options are considered. First, a local search is executed. Outstanding replenishment (or repair) orders are checked for possible expedition, or stocking locations in the neighborhood (possibly in adjacent countries) are checked. Second, if the local search attempt does not result in a feasible solution, a central search is executed. All stock in the Service Part Supply System (located at the central warehouse or downstream in the distribution network) is checked, including pipeline stock. If necessary, cannibalization of equipment is applied or the part is bought at a local vendor. If this central search does not result in a satisfying solution, a third and final option is executed. Outstanding orders at the manufacturer (or vendor) are investigated or an emergency replenishment order directly from the supplier is issued.

5.8.2 Parameter setting

The normal replenishment cost and lead times are summarized in table 5.8. The lead time is expressed in days and exists of order planning time plus transportation time. The cost of a normal replenishment as shown in table 5.8 is derived from the real value by normalizing the central replenishment cost. In our model we assume that the manufacturing plant is located in Spain (which in practice represents the largest European production plant). Since the normal replenishment cost from the plant to the central warehouse is not known, we assume that this cost is equal to the cost of a normal replenishment from the central warehouse to the national warehouse in Spain.

| | Central | National Warehouses | | | | | | |
|-----------|-----------|---------------------|------|------|------|------|------|------|
| | Warehouse | Fi | Fr | Ge | No | Sp | Sw | UK |
| LEAD TIME | 49 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| COST | 1 | 1.23 | 0.35 | 0.27 | 1.02 | 1.00 | 0.89 | 0.73 |

Table 5.8: Normal replenishment lead times (in days) and cost (normalized values)

Emergency Supply Model

The cost and lead times of the emergency replenishment structure are presented in Appendix E. For both the emergency lateral transshipments between national warehouses and the direct shipments from the central warehouse we can choose between two options: delivery within 16 hours or delivery within 68 hours. The fast option is of course more expensive than the slow option. The cost of using these lateral and direct deliveries is based on a fixed cost per shipment plus a variable cost depending on the weight of the service part. The direct shipment lead time from the factory is either 40 hours (24 hours planning lead time plus 16 hours delivery time) or 92 hours (24 hours planning lead time plus 68 hours delivery time). The cost of these direct shipments from the plant in Spain is based on the emergency lateral transshipment cost between Spain and the national warehouse under consideration.

Since the company used an inventory holding cost percentage of 41% per year for planning purposes, we used this percentage as well in our model calculations. Finally, the penalty cost per day when customers have to wait for backordered service parts is set equal to an equivalent of 1000 Dutch guilders per day.

5.8.3 Numerical results

For our experiment we selected eight pilot service parts with different prices, ranging from 100 to 100.000 Dutch guilders. The characteristics of these pilot parts are summarized in table 5.9. The national stock levels for the different parts are determined by summing up the stock levels of all inventory locations within that specific country. Consequently, the national stock level may seem very high in certain cases, while in reality it represents different inventory locations within one country. Using the data from table 5.9 in our Emergency Supply Model, we calculated the service performance and cost performance for each service part individually. The service performance at the national warehouses is expressed by the fractions α (fraction of total demand satisfied through lateral transshipments) and β (fraction of total demand satisfied from stock on hand). The fractions γ (fraction of total demand satisfied through direct shipments from the central warehouse) and θ (fraction of total demand satisfied through direct shipments from the factory) are not shown since they were equal to zero for seven of the eight parts. Only for part 6 the θ -value was slightly positive ($\theta=0.009$). The cost performance is expressed by the expected total daily cost of operating the Service Part Supply System. The total cost consists of four cost factors: inventory holding cost at all inventory locations, normal replenishment cost for all inventory locations with positive stock levels, emergency replenishment cost for the local warehouses in case of a demand in stock-out situations, and penalty cost when customers have to wait for backordered service parts. Tables 5.10 and 5.11 show the service and cost performance of the eight service parts when the fast (and expensive) emergency supply structure is applied. The service performance is only shown for those parts with positive demand at the various national warehouses.

| PART | PRICE | WEIGHT | S_{cw} | (DEMAND, STOCK LEVEL) | | | | | | |
|------|-------|--------|----------|-----------------------|-------|--------|-------|-------|-------|--------|
| | | | | Fi | Fr | Ge | No | Sp | Sw | UK |
| 1 | 100 | 5.10 | 3 | (1,1) | (3,6) | (1,5) | (1,1) | (0,1) | (0,4) | (1,7) |
| 2 | 100 | 0.25 | 3 | (0,0) | (1,1) | (6,2) | (0,0) | (0,0) | (0,1) | (2,2) |
| 3 | 1004 | 0.23 | 20 | (0,0) | (8,4) | (9,3) | (1,1) | (2,1) | (2,0) | (10,3) |
| 4 | 1010 | 3.40 | 3 | (0,0) | (5,1) | (5,5) | (3,1) | (0,0) | (5,1) | (7,1) |
| 5 | 9961 | 2.00 | 3 | (0,0) | (6,1) | (11,3) | (1,1) | (9,3) | (0,1) | (3,2) |
| 6 | 10760 | 3.30 | 3 | (0,0) | (2,0) | (10,3) | (0,1) | (3,0) | (0,0) | (0,0) |
| 7 | 98057 | 65.0 | 3 | (0,0) | (0,1) | (0,2) | (1,1) | (0,0) | (0,0) | (0,0) |
| 8 | 95165 | 35.0 | 3 | (0,0) | (0,1) | (3,1) | (0,1) | (0,1) | (0,0) | (0,0) |

Table 5.9: Characteristics of the eight pilot parts (price in Dutch guilders, weight in kilograms, central stock level, annual demand and stock levels for all national warehouses)

| PART | SERVICE PERFORMANCE | | | | | | | |
|------|---------------------|-------|-------|-------|-------|-------|-------|-------|
| | Fi | Fr | Ge | No | Sp | Sw | UK | |
| 1 | α : | 0.027 | 0.000 | 0.000 | 0.027 | - | - | 0.000 |
| | β : | 0.973 | 1.000 | 1.000 | 0.973 | - | - | 1.000 |
| 2 | α : | - | 0.029 | 0.013 | - | - | - | 0.002 |
| | β : | - | 0.971 | 0.987 | - | - | - | 0.998 |
| 3 | α : | - | 0.000 | 0.002 | 0.024 | 0.047 | 1.000 | 0.002 |
| | β : | - | 1.000 | 0.998 | 0.976 | 0.953 | 0.000 | 0.998 |
| 4 | α : | - | 0.234 | 0.000 | 0.159 | - | 0.239 | 0.300 |
| | β : | - | 0.766 | 1.000 | 0.841 | - | 0.761 | 0.700 |
| 5 | α : | - | 0.296 | 0.036 | 0.066 | 0.022 | - | 0.018 |
| | β : | - | 0.704 | 0.964 | 0.934 | 0.978 | - | 0.982 |
| 6 | α : | - | 0.991 | 0.003 | - | 0.991 | - | - |
| | β : | - | 0.000 | 0.988 | - | 0.000 | - | - |
| 7 | α : | - | - | - | 0.024 | - | - | - |
| | β : | - | - | - | 0.976 | - | - | - |
| 8 | α : | - | - | 0.070 | - | - | - | - |
| | β : | - | - | 0.930 | - | - | - | - |

Table 5.10: Service performance when the fast emergency supply structure is applied

| PART | COST PERFORMANCE | | | | |
|------|------------------|------|------|------|--------|
| | ihc | nrc | erc | pc | tc |
| 1 | 3.02 | 0.14 | 0.02 | 0.10 | 3.27 |
| 2 | 0.85 | 0.01 | 0.01 | 0.20 | 1.07 |
| 3 | 30.41 | 0.03 | 0.28 | 3.93 | 34.66 |
| 4 | 9.38 | 0.33 | 1.08 | 9.09 | 19.88 |
| 5 | 105.27 | 0.23 | 0.35 | 4.59 | 110.43 |
| 6 | 56.05 | 0.17 | 0.87 | 9.72 | 66.80 |
| 7 | 753.52 | 0.32 | 0.03 | 0.04 | 753.91 |
| 8 | 697.88 | 0.32 | 0.12 | 0.38 | 698.70 |

Table 5.11: Cost performance when the fast emergency supply structure is applied (ihc:inventory holding cost,nrc:normal replenishment cost,erc:emergency replenishment cost, pc: penalty cost, tc:total cost)

We see that the fill rates (β) are in general very high (i.e. > 90%) at the national warehouses with positive stock levels, with the exception of part 4 (at the national warehouses in France, Norway, Sweden, and the UK) and part 5 (at the national warehouse in France). The stock levels for most parts are high enough to realize a high fill rate performance. As a result, the inventory holding cost is the dominant cost factor for these parts. This is illustrated in table 5.12. For every part we calculated the average fill rate (weighed by demand), the average lateral transshipment rate (weighed by demand), and the percentage of the total cost that is caused by the inventory holding cost factor.

| PART | $\sum_i (\lambda_i / \lambda_0) \beta_i$ | $\sum_i (\lambda_i / \lambda_0) \alpha_i$ | (ihc/tc)*100% |
|------|--|---|---------------|
| 1 | 0.992 | 0.008 | 92.4 |
| 2 | 0.988 | 0.012 | 79.4 |
| 3 | 0.933 | 0.067 | 87.7 |
| 4 | 0.802 | 0.198 | 47.2 |
| 5 | 0.917 | 0.083 | 95.3 |
| 6 | 0.659 | 0.341 | 83.9 |
| 7 | 0.976 | 0.024 | 99.9 |
| 8 | 0.930 | 0.070 | 99.9 |

Table 5.12: Weighed service performance and holding cost percentage for all parts

The percentage of the total cost caused by inventory holding cost is for all parts very high, except for part 4 which has a relatively low fill rate (i.e. $\beta=0.802$). For this part the penalty cost percentage is high as well, since relatively many customers have to wait for a lateral transshipment. However, part 6 has an even lower fill rate (i.e. $\beta=0.659$) but still a high holding cost percentage. In absolute figures, the penalty cost is very high, but since part 6 is ten times more expensive than part 4, the relative importance of the penalty cost is low compared to the inventory holding cost.

The normal replenishment cost and the emergency replenishment cost are for all parts relatively low, regardless of service performance and part characteristics. We calculated the service and cost performance for these eight parts as well when applying the slow (and relatively cheap) emergency supply structure. The service performance is identical to the results displayed in table 5.10 and is therefore not presented. The cost performance of the alternative supply structure is presented in table 5.13.

| PART | EMERGENCY SUPPLY STRUCTURE | | | | | |
|------|----------------------------|------|--------|------|-------|--------|
| | FAST | | | SLOW | | |
| | erc | pc | tc | erc | pc | tc |
| 1 | 0.02 | 0.10 | 3.27 | 0.01 | 0.41 | 3.58 |
| 2 | 0.01 | 0.20 | 1.07 | 0.01 | 0.86 | 1.73 |
| 3 | 0.28 | 3.93 | 34.66 | 0.28 | 16.70 | 47.42 |
| 4 | 1.08 | 9.09 | 19.88 | 0.88 | 38.48 | 49.07 |
| 5 | 0.35 | 4.59 | 110.43 | 0.28 | 19.40 | 125.17 |
| 6 | 0.87 | 9.72 | 66.80 | 0.65 | 40.11 | 96.98 |
| 7 | 0.03 | 0.04 | 753.91 | 0.02 | 0.19 | 754.04 |
| 8 | 0.12 | 0.38 | 698.70 | 0.04 | 1.62 | 699.87 |

Table 5.13: Cost performance when the slow emergency supply structure is applied

Only the emergency replenishment cost (erc) and the penalty cost (pc) are affected when using the slow emergency supply structure. It is obvious from the results in the table that the (marginal) reduction in emergency replenishment cost does not compensate the sometimes enormous increase in penalty cost. Part 4, for example, shows an increase of 147% in total cost when using the slow emergency supply structure. For the expensive parts (7 and 8) the increase in penalty cost is not significant, due to the high inventory holding cost.

5.8.4 Optimization results

In the previous section we calculated the service and cost performance of the Service Part Supply System for given stock levels of the eight pilot parts. These stock levels are not optimal from an integral cost point of view. One of the reasons is that the national stock levels are in fact the sum of regional stock levels within one country. Due to customer response time restrictions it may be necessary to stock service parts at warehouses in different regions in one country, although the model might advise to stock only one service part in that country. Since our model is a two-echelon model, it can not account for regional differences with respect to e.g. response time requirements. Another reason that the practical stock levels are not cost-optimal is the competence of the national organizations to decide how many service parts should be stocked at the individual national warehouses. The national organizations are responsible for their own win and loss situation and therefore have the authority to decide on the height of the national stock levels of service parts. As a result of this organizational structure, the individual decisions by the national organizations may conflict with the overall cost performance of the Service Part Supply System.

Keeping in mind that there are good reasons to use stock levels in practice that deviate from the optimal values according to our Emergency Supply Model, we calculated these optimal stock levels for the eight pilot parts. We applied enumeration to calculate the optimal stock levels and the results are shown in figure 5.14. The results show that major cost reductions can be obtained when using these cost-optimal stock levels. However, two remarks have to be made. First, the demand data we used refers to the demand for parts in the period April 1995 until March 1996. If there was no demand for a specific service part in this period at a specific national warehouse, the model assumes that there is never a demand at this national warehouse for this part. Consequently, no stock is ever allocated to this warehouse. Second, the optimal stock levels are calculated assuming a penalty cost of one thousand Dutch guilders per day. The penalty cost for the expensive parts 7 and 8, however, are likely to be much higher than for the cheaper parts. A penalty cost of e.g. one thousand Dutch guilders *per hour* is very realistic for these parts. In that case the optimal stock levels for parts 7 and 8 are equal to one at the national warehouses with a positive demand and zero at the central warehouse. The cost savings are then 85.3% for part 7 and 74.4% for part 8.

| PART | S _{cw} | S _{Fi} | S _{Fr} | S _{Ge} | S _{No} | S _{Sp} | S _{Sw} | S _{Uk} | TC | % |
|------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-------|------|
| 1 | 2 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 1.25 | 61.8 |
| 2 | 2 | 0 | 1 | 3 | 0 | 0 | 0 | 2 | 0.87 | 18.7 |
| 3 | 3 | 0 | 3 | 3 | 1 | 1 | 1 | 3 | 13.88 | 60.0 |
| 4 | 4 | 0 | 2 | 2 | 1 | 0 | 2 | 2 | 12.37 | 37.8 |
| 5 | 2 | 0 | 1 | 2 | 0 | 2 | 0 | 1 | 63.92 | 42.1 |
| 6 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 39.64 | 40.7 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5.85 | 99.2 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 14.82 | 97.9 |

Table 5.14: Optimal stock levels at the central warehouse (cw) and the countries and cost performance (TC: total cost; %: cost reduction as percentage of cost in table 5.11)

5.8.5 Economic trade-offs

The results presented in the previous sections are all very much dependent on the parameter setting in the supply system. We assumed for example a fixed inventory holding cost percentage and a fixed penalty cost per day for backordered service parts. Changing these values may affect the performance of the supply system significantly. Investigating the effect of changing the values of different parameters on the cost and service performance of the system is a topic of research on its' own. The Emergency Supply Model presented in this chapter is a valuable tool for this line of research. The outcome of the model should not be interpreted as 'the one and only' optimal solution to a practical problem. However, the model does give a better understanding of the relationships between different parameters in the supply system and offers the possibility to consider economic trade-offs.

5.9 Conclusions

In this chapter we presented and analyzed the Emergency Supply Model: a two-echelon inventory system with supply flexibility. Customers that arrive at the local warehouses in a stock-out situation are not backordered but satisfied through an emergency lateral transshipment, a direct delivery from the central warehouse, or a direct delivery from the plant. The Emergency Supply Model is an approximate model that is solved in two steps. In the first step an exact aggregate model is developed that combines all local warehouses into one warehouse. In the second step we use an approximation technique to calculate the service performance at the various local warehouses. The numerical results indicate that the performance of our model is very close to the simulation results. Another important observation is that the distribution of the shipment times has a negligible impact

Emergency Supply Model

on the service performance. This is important since we have to assume exponentially distributed shipment times in our model analysis. In an economical analysis we compared the optimal cost results of our model with the VARI-METRIC cost results. In all cases we found major cost reductions, which indicate that using emergency supply flexibility in a distribution network for service parts can be very beneficial.

One of the assumptions in the Emergency Supply Model is the use of complete pooling: every local warehouse is checked for excess stock before a direct delivery from the central warehouse is made. We compared this complete pooling concept with two alternative pooling concepts: fixed pooling (a local warehouse in a stock-out situation can make a request for a lateral transshipment at all other local warehouses that are part of its own fixed pooling group) and variable pooling (a local warehouse in a stock-out situation can make a request for a lateral transshipment at its two neighboring local warehouses). The numerical results showed that the differences in service performance under the fixed pooling concept and under the variable pooling concept is negligible. The differences with the complete pooling concept are only significant for low fill rates, i.e. lower than 70 %. This means that the pooling structure is only important when the local warehouses are understocked, i.e. give a low fill rate performance. When supply flexibility is viewed as a back-up facility in case of occasional stock-out situations at local warehouses, implying that most of the demand is satisfied directly from stock on hand, the type of pooling structure (complete, fixed, or variable) that is used for lateral shipments does not affect the service performance of the system significantly.

The Emergency Supply Model can be used to investigate economic trade-offs. In a case study we showed how the service and cost performance of a service part supply system are related to parameters like the cost price of a service part, the cost of using emergency shipments, and penalties that must be paid when customers have to wait for backordered service parts. The model is an excellent tool for providing insight into the complex relations between parameter values in a Service Part Supply System.

In this chapter we compared the performance of the Emergency Supply Model with the performance of the VARI-METRIC model in which no flexibility is applied at all. The Emergency Supply Model prescribes how to react in case of a stock-out situation. Comparison of this policy with other policies in which only a limited number of supply flexibility options is available (or other types of flexibility are applied) is needed to investigate the added value of individual supply flexibility options. In the next chapter we evaluate different policies for applying flexibility by means of simulation.

Chapter 6

Policy evaluation

6.1 Introduction

In Chapters 4 and 5 we presented two analytical models (i.e., the Emergency Repair Model and the Emergency Supply Model) that can be used to evaluate the effectiveness of two specific flexibility options: 1) the use of emergency repair (or supply) for an inventory location at the lowest level in the network, and 2) the use of emergency lateral transshipments and direct deliveries between inventory locations in a two-echelon inventory system. The policy applied in the latter model (see Chapter 5) prescribes that the demand that cannot be satisfied from stock on hand, is satisfied by checking a number of emergency supply options in a fixed order: first lateral transshipment, second direct delivery from the central warehouse, and third direct delivery from the plant (which can always be applied). In this chapter we evaluate *different* policies that can be applied when demand cannot be satisfied from stock on hand. The policies that are considered differ from each other with respect to the order of events that is prescribed in case of a stock-out situation at an inventory location that serves external customers. The evaluation of the different policies is based on a cost analysis and service performance analysis. Modeling and analyzing the characteristics of such policies is very complex, if not impossible. We therefore use simulation to evaluate the performance of the various flexibility policies.

In Section 6.2 we present the distribution network configuration that was used to evaluate the different policies. The performance measures (cost and service measures) that are used for the evaluation are also introduced in this section. In Section 6.3 we discuss the parameter settings for the simulation experiment. In Section 6.4 twelve different policies are presented that prescribe a different order of emergency replenishment events that have to be carried out in case of a stock-out situation. In Section 6.5 we discuss experimental design issues. The numerical evaluation of the twelve policies is given in Section 6.6. Finally, in Section 6.7 we present some general conclusions with respect to the outcome of the simulation experiment.

6.2 The distribution network configuration

We consider a three-echelon network of inventory locations consisting of a central warehouse, three identical national warehouses, and nine identical local warehouses. The central warehouse is supplied by a production plant which is assumed to have infinite capacity. The network is visualized in figure 6.1.

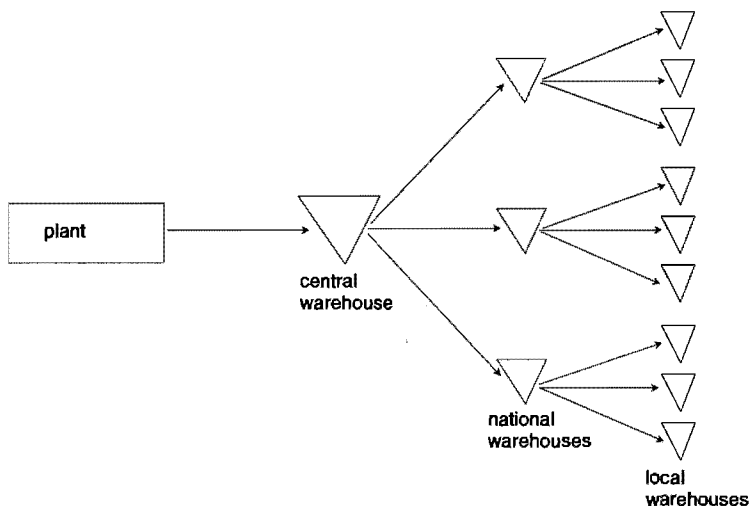


Figure 6.1: The network model

The following assumptions are made:

- 1) Customer demand for service parts originates at the local warehouses only, and follows a Poisson process.
- 2) All lead times between inventory locations are deterministic.
- 3) All inventory locations in the network apply an (S-1, S) ordering policy.
- 4) All local warehouses are identical with respect to demand rate and lead time.

The performance of the network is measured in terms of service and cost (see also the previous chapters). The service performance with respect to external customers that arrive at the local warehouses, is measured in two ways:

Fill rate : fraction of demand at a local warehouse that is satisfied from stock on hand
Response time : average waiting time for a customer arriving in a stock-out situation at an arbitrary local warehouse

The service performance is defined for an arbitrary local warehouse since all local warehouses are assumed to be identical. The total cost of operating the network is defined as the sum of the following four cost factors:

Policy evaluation

- 1) Inventory holding cost at all inventory locations in the network
- 2) Normal replenishment cost for all inventory locations in the network
- 3) Emergency replenishment cost for all local warehouses in the network
- 4) Penalty cost for all customers who must wait at a local warehouse in the network

These three performance indicators (fill rate, average response time, and total cost) are measured in the simulation experiment for different configurations of the distribution network under consideration. In the next section we describe the parameters that determine the network configuration.

6.3 Parameter settings

Many parameters determine the structure of a distribution network such as the one presented in figure 6.1. In order to limit the size of the simulation experiment we restrict ourselves to the following five parameters to characterize the distribution network:

- λ : daily demand rate for service parts at a local warehouse
- C : purchase price of one service part
- E : emergency replenishment structure expressed in time and cost
- h : inventory holding cost at an arbitrary warehouse
- p : penalty cost per day for customers who have to wait for a backordered service part

We believe that these five parameters are very important for characterizing the Service Part Supply System. Other parameters (such as normal replenishment lead times and normal replenishment cost) are fixed in the simulation experiment. In practice these (normal replenishment) parameters are difficult to influence since they concern routine activities. Normal replenishment cost, for example, is usually based on standard contracts with service providers who take care of the transportation of goods. In general, service parts make up a minor part of the total amount of goods that have to be transported. The service parts just "go along the ride" with the other goods. Normal replenishment lead times are also fixed in the simulation experiment. It is also assumed that these lead times are deterministic. This choice is supported by the insensitivity results we found in Chapters 4 and 5 with respect to the choice of lead time distribution.

Each of the five network parameters can take a low (symbol: -) and a high (symbol: +) value. The daily demand rate is 0.01 (representing a slow moving part with three to four demands per year) or 0.05 (representing a 'fast moving' part with fifteen to twenty demands per year). The cost price of one service part is either one thousand Dutch guilders (representing a 'cheap' service part) or one

hundred thousand Dutch guilders (representing an expensive service part). The inventory holding cost per item per year is expressed as a percentage of the unit price; it is either 10 % or 25 %. The factors that determine this percentage are interest rate, insurance rate, and risk of obsolescence. The daily penalty cost is one thousand Dutch guilders (representing less critical service parts with minor consequences in case of a failure) or one hundred thousand Dutch guilders (representing vital service parts with severe consequences in case of failure). Table 6.1 gives an overview of the two extreme values of daily demand rate (λ), the unit price (C), the inventory holding cost (h), and the daily penalty cost (p).

| parameter | - | + |
|-----------|-----------|-------------|
| λ | 0.01 | 0.05 |
| C | DFL 1.000 | DFL 100.000 |
| h | 10 % | 25 % |
| p | DFL 1.000 | DFL 100.000 |

Table 6.1: Extreme values for the network parameters

Finally, the emergency replenishment structure is either "slow and cheap" (symbol: -) or "fast and expensive" (symbol: +). The associated numerical values (expressed in time and cost) are given in table 6.2. These values are obtained from the case study described in Section 5.8. The emergency replenishment structure exists of four elements: 1) Emergency Lateral Transhipments (ELT) between local warehouses in the same country; 2) Direct Deliveries from the National Warehouses (DD-N); 3) Direct Deliveries from the Central Warehouse (DD-C); and 4) Direct Deliveries from the Factory (DD-F). For example, the fast and expensive option for Emergency Lateral Transhipments corresponds with a delivery time of 0.17 days (or 4 hours) and a price of DFL 11. The '+' values for the emergency replenishment structure are obtained from the '-' values by reducing the lead times with 50% and increasing the cost with 10%.

| value | ELT | DD-N | DD-C | DD-F |
|-------|-----------|-----------|--------|--------|
| - | (0.33;10) | (0.67;20) | (2;40) | (4;60) |
| + | (0.17;11) | (0.33;22) | (1;44) | (2;66) |

Table 6.2: Extreme values for the emergency replenishment structure;
(x;y) : x = lead time in days and y = cost of one shipment in Dutch guilders

In the experiment the normal replenishment lead times are deterministic; they are equal to 30 days for the central warehouse, 10 days for the national warehouses, and 5 days for the local warehouses. These lead times consist of transportation time and order planning time; the latter (administrative order planning time) usually makes up the larger part of the lead time. The cost of normal replenishments throughout the network is set equal to one Dutch guilder per shipment. These data correspond with the data observed in practice (see Section 5.8).

The stock levels at the various inventory locations in the SPSS that are used as input for the simulation model are derived from an approximate model by Muckstadt and Thomas (1980); see also Chapter 2. They model the use of direct deliveries from inventory locations at a higher echelon in a two-echelon system, in case of a stock-out at an inventory location at the lowest echelon. They do not model the use of emergency lateral transshipments or checking the pipeline for outstanding orders. We extend their model to a three-echelon system, and calculate the minimal cost stock levels at all inventory locations. These stock levels are used as input to the simulation model for evaluating the different flexibility policies. Therefore this specific stock allocation, derived from the analytic model, does not necessarily have to be optimal for the various policies.

6.4 Policy descriptions

The goal of the simulation experiment is to evaluate different emergency replenishment policies for customers who arrive in a stock-out situation at a local warehouse. In total we evaluate twelve emergency replenishment policies. Policies 1 through 6 are described below. Policies 7 through 12 are identical to policies 1 through 6 with the extra flexibility option of checking the pipeline for outstanding orders in case of a stock-out situation. This flexibility option prevents the use of other, probably more expensive, emergency replenishment procedures (i.e. ELT, DD-N, DD-C, or DD-F) with lead times that exceed the remaining lead time of a normal replenishment order that is in the pipeline. Only the pipeline between the national warehouse and the local warehouse under consideration is checked! The following policies are implemented in the simulation model.

Policy 1: If demand for a service part cannot be satisfied from stock on hand at a local warehouse, a direct delivery from the factory is applied.

Policy 2: If demand for a service part cannot be satisfied from stock on hand at a local warehouse, an emergency lateral transshipment from an arbitrary local warehouse (that has the same national warehouse) is applied. Otherwise a direct delivery from the factory is applied.

- Policy 3:** If demand for a service part cannot be satisfied from stock on hand at a local warehouse, a direct delivery from the national warehouse is applied. Otherwise a direct delivery from the factory is applied.
- Policy 4:** If demand for a service part cannot be satisfied from stock on hand at a local warehouse, a direct delivery from the national warehouse is applied. If this is not possible, a direct delivery from the central warehouse is applied. Otherwise a direct delivery from the factory is applied.
- Policy 5:** If demand for a service part cannot be satisfied from stock on hand at a local warehouse, an emergency lateral transshipment from an arbitrary local warehouse (that has the same national warehouse) is applied. If this is not possible, a direct delivery from the national warehouse is applied. Otherwise a direct delivery from the factory is applied.
- Policy 6:** If demand for a service part cannot be satisfied from stock on hand at a local warehouse, an emergency lateral transshipment from an arbitrary local warehouse (that has the same national warehouse) is applied. If this is not possible, a direct delivery from the national warehouse or central warehouse is applied. Otherwise a direct delivery from the factory is applied.

Table 6.3 summarizes the emergency replenishment options that are associated with each of the twelve policies.

| Option | Policy | | | | | | | | | | | |
|--------|--------|---|---|---|---|---|---|---|---|----|----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| P | | | | | | | * | * | * | * | * | * |
| ELT | | * | | | * | * | | * | | | * | * |
| DD-N | | | * | * | * | * | | | * | * | * | * |
| DD-C | | | | * | | * | | | | * | | * |
| DD-F | * | * | * | * | * | * | * | * | * | * | * | * |

Table 6.3: P = pipeline
 ELT = emergency lateral transshipment
 DD-N = direct delivery from national warehouse
 DD-C = direct delivery from central warehouse
 DD-F = direct delivery from factory

6.5 Design issues

The configuration of the distribution network is determined by five parameters that are varied over two levels. Therefore we have $2^5=32$ extreme network configurations. In addition we have 12 emergency replenishment policies for each network configuration. If we want to do a full factorial design experiment, we have to simulate 384 combinations of network configurations and policies. In order to limit the number of simulations, we apply a fractional factorial design (Kleijnen and Van Groenendaal, 1992, page 175) in which we consider $2^{5-2} = 8$ network configurations. The five network parameters in this design are varied as follows:

| Case | λ | C | E | h (= λC) | p (= λC) |
|------|-----------|---|---|-----------------------|-----------------------|
| 1 | + | + | + | + | + |
| 2 | - | + | + | - | - |
| 3 | + | - | + | - | + |
| 4 | - | - | + | + | - |
| 5 | + | + | - | + | - |
| 6 | - | + | - | - | + |
| 7 | + | - | - | - | - |
| 8 | - | - | - | + | + |

Table 6.4: fractional factorial 2^{5-2} design

Combining these 8 scenarios and 12 policies results in 96 combinations. Every simulation run consists of an initialization phase and ten subruns. The length of the initialization phase was determined graphically, according to Welch (Law and Kelton, 1991, page 546). The length of the subruns was determined by using the Von Neumann ratio (Kleijnen and Van Groenendaal, 1992, page 192). More details about the application of these methods to our experiment can be found in Appendix F. The length of the initialization phase equals the arrival of approximately 2000 customers per local warehouse. The length of every subrun equals the arrival of approximately 1000 customers per local warehouse.

6.6 Numerical results

The first step is to calculate the stock levels for the different inventory locations in the distribution network. As mentioned before, we used the analytical model described by Muckstadt and Thomas (1980) to determine the stock levels that minimize the total cost. The stock levels that minimize the

total cost function for the eight cases under consideration are presented in table 6.5. The sum of the stock levels of all inventory locations in the distribution system (i.e., system stock = $S_{cw} + 3*S_{nw} + 9*S_{lw}$) is also shown. We used these stock levels, together with the appropriate choice of network parameters and policies, to fix the simulation experiment.

| Case | S_{cw} | S_{nw} | S_{lw} | system stock |
|------|----------|----------|----------|--------------|
| 1 | 12 | 3 | 3 | 48 |
| 2 | 2 | 1 | 0 | 5 |
| 3 | 11 | 4 | 5 | 68 |
| 4 | 3 | 2 | 1 | 18 |
| 5 | 10 | 3 | 1 | 28 |
| 6 | 3 | 1 | 2 | 24 |
| 7 | 13 | 4 | 3 | 52 |
| 8 | 3 | 1 | 3 | 33 |

Table 6.5: Stock levels for the central warehouse (S_{cw}), national warehouses (S_{nw}), and local warehouses (S_{lw}) according to the Muckstadt-Thomas model

The simulation results with respect to the performance indicators are presented in tables 6.6, 6.7, and 6.8. Table 6.6 shows the average total cost and the associated 95%-confidence interval for each of the 96 combinations of network configurations and emergency replenishment policies. The corresponding average response time and fill rate results are presented in tables 6.7 and 6.8. In the remainder of this section we discuss four topics in relation to these numerical results. In Section 6.6.1 we investigate the added value of including the pipeline flexibility option in the emergency replenishment structure. In Section 6.6.2 we identify and discuss the optimal policies for the cases under consideration. In Section 6.6.3 we investigate the main effects of the five network parameters. Finally, in Section 6.6.4 we focus our attention on one specific network parameter: the emergency replenishment structure.

| Case | Policy | | | | | | | | | | | |
|------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 2748 ±60.4 | 1944 ±14.6 | 2744 ±57.5 | 2731 ±55.8 | 1944 ±14.6 | 1944 ±14.6 | 2644 ±53.1 | 1941 ±15.8 | 2642 ±51.9 | 2637 ±50.6 | 1941 ±15.8 | 1941 ±15.8 |
| 2 | 186 ±1.1 | 186 ±1.1 | 94 ±2.3 | 91 ±2.3 | 94 ±2.3 | 91 ±2.3 | 186 ±1.1 | 186 ±1.1 | 94 ±2.3 | 91 ±2.3 | 94 ±2.3 | 91 ±2.3 |
| 3 | 44 ±10.3 | 17 ±0.9 | 44 ±10.3 | 44 ±10.3 | 17 ±0.9 | 17 ±0.9 | 39 ±10.6 | 17 ±0.9 | 39 ±10.6 | 39 ±10.6 | 17 ±0.9 | 17 ±0.9 |
| 4 | 18 ±0.4 | 9 ±0.1 | 13 ±0.2 | 13 ±0.2 | 9 ±0.1 | 9 ±0.1 | 16 ±0.3 | 9 ±0.1 | 13 ±0.2 | 12 ±0.2 | 9 ±0.1 | 9 ±0.1 |
| 5 | 1144 ±7.8 | 740 ±6.2 | 952 ±7.3 | 944 ±7.4 | 738 ±5.8 | 735 ±6.0 | 994 ±8.1 | 723 ±5.8 | 915 ±7.1 | 910 ±6.9 | 722 ±6.1 | 721 ±5.9 |
| 6 | 674 ±15.9 | 510 ±2.8 | 674 ±16.1 | 671 ±16.3 | 510 ±2.8 | 510 ±2.8 | 630 ±12.8 | 509 ±2.7 | 630 ±12.8 | 628 ±12.7 | 509 ±2.7 | 509 ±2.7 |
| 7 | 19 ±0.8 | 12 ±0.2 | 19 ±0.8 | 19 ±0.8 | 12 ±0.2 | 12 ±0.2 | 16 ±0.7 | 11 ±0.2 | 16 ±0.7 | 16 ±0.7 | 11 ±0.2 | 11 ±0.2 |
| 8 | 30 ±3.6 | 20 ±0.3 | 30 ±3.6 | 30 ±3.6 | 20 ±0.3 | 20 ±0.3 | 27 ±3.1 | 20 ±0.3 | 27 ±3.1 | 27 ±3.1 | 20 ±0.3 | 20 ±0.3 |

Table 6.6: Average total cost and individual 95% confidence interval

| Case | Policy | | | | | | | | | | | |
|------|--------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 2.00 | 0.22 | 1.99 | 1.96 | 0.22 | 0.22 | 1.58 | 0.21 | 1.58 | 1.57 | 0.21 | 0.21 |
| | - | ± 0.02 | ± 0.01 | ± 0.02 | ± 0.02 | ± 0.02 | ± 0.03 | ± 0.02 | ± 0.04 | ± 0.04 | ± 0.02 | ± 0.02 |
| 2 | 2.00 | 2.00 | 0.86 | 0.83 | 0.86 | 0.83 | 2.00 | 2.00 | 0.86 | 0.83 | 0.86 | 0.83 |
| | - | - | ± 0.01 | ± 0.01 | ± 0.01 | ± 0.01 | - | - | ± 0.01 | ± 0.01 | ± 0.01 | ± 0.01 |
| 3 | 2.00 | 0.17 | 2.00 | 2.00 | 0.17 | 0.17 | 1.62 | 0.16 | 1.62 | 1.62 | 0.16 | 0.16 |
| | - | ± 0.00 | - | - | ± 0.00 | ± 0.00 | ± 0.25 | ± 0.01 | ± 0.25 | ± 0.25 | ± 0.01 | ± 0.01 |
| 4 | 2.00 | 0.21 | 0.96 | 0.94 | 0.21 | 0.21 | 1.64 | 0.21 | 0.86 | 0.84 | 0.21 | 0.21 |
| | - | ± 0.01 | ± 0.02 | ± 0.02 | ± 0.01 | ± 0.01 | ± 0.02 | ± 0.01 | ± 0.02 | ± 0.02 | ± 0.01 | ± 0.01 |
| 5 | 4.00 | 1.28 | 2.65 | 2.61 | 1.28 | 1.26 | 2.89 | 1.23 | 2.34 | 2.32 | 1.22 | 1.22 |
| | - | ± 0.02 | ± 0.03 | ± 0.02 | ± 0.02 | ± 0.02 | ± 0.02 | ± 0.02 | ± 0.03 | ± 0.02 | ± 0.02 | ± 0.02 |
| 6 | 4.00 | 0.35 | 3.99 | 3.91 | 0.35 | 0.35 | 2.88 | 0.34 | 2.87 | 2.86 | 0.34 | 0.34 |
| | - | ± 0.03 | ± 0.01 | ± 0.06 | ± 0.03 | ± 0.03 | ± 0.12 | ± 0.03 | ± 0.12 | ± 0.13 | ± 0.03 | ± 0.03 |
| 7 | 4.00 | 0.38 | 3.64 | 3.58 | 0.38 | 0.38 | 2.14 | 0.36 | 2.06 | 2.05 | 0.36 | 0.36 |
| | - | ± 0.03 | ± 0.12 | ± 0.11 | ± 0.03 | ± 0.03 | ± 0.11 | ± 0.03 | ± 0.14 | ± 0.14 | ± 0.03 | ± 0.03 |
| 8 | 4.00 | 0.33 | 4.00 | 4.00 | 0.33 | 0.33 | 2.90 | 0.31 | 2.90 | 2.90 | 0.31 | 0.31 |
| | - | - | - | - | - | - | ± 0.59 | ± 0.03 | ± 0.59 | ± 0.59 | ± 0.03 | ± 0.03 |

Table 6.7: Average response time and individual 95% confidence interval

| Case | Policy | | | | | | | | | | | |
|------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | 99.0 ±0.1 | 98.8 ±0.1 | 99.0 ±0.1 | 99.0 ±0.1 | 98.8 ±0.1 | 98.8 ±0.1 | 98.9 ±0.1 | 98.8 ±0.1 | 98.9 ±0.1 | 98.9 ±0.1 | 98.8 ±0.1 | 98.8 ±0.1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 |
| 4 | 94.9 ±0.2 | 94.4 ±0.2 | 94.9 ±0.2 | 94.9 ±0.2 | 94.4 ±0.2 | 94.4 ±0.2 | 94.8 ±0.2 | 94.3 ±0.2 | 94.8 ±0.2 | 94.8 ±0.2 | 94.3 ±0.2 | 94.3 ±0.2 |
| 5 | 78.1 ±0.3 | 67.2 ±0.5 | 76.1 ±0.3 | 76.0 ±0.3 | 67.2 ±0.5 | 67.1 ±0.5 | 72.2 ±0.4 | 65.1 ±0.5 | 72.3 ±0.4 | 72.3 ±0.4 | 65.2 ±0.5 | 65.2 ±0.6 |
| 6 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 | 99.5 ±0.0 |
| 7 | 99.5 ±0.1 | 99.5 ±0.1 | 99.5 ±0.1 | 99.5 ±0.1 | 99.5 ±0.1 | 99.5 ±0.1 | 99.5 ±0.1 | 99.4 ±0.1 | 99.5 ±0.1 | 99.5 ±0.1 | 99.4 ±0.1 | 99.4 ±0.1 |
| 8 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 | 100.0 ±0.0 |

Table 6.8: Average fill rate and individual 95% confidence interval

6.6.1 Pipeline flexibility

The simulation results enable us to investigate the added value (in terms of total cost, average response time, and fill rate) of checking the pipeline for outstanding orders in case of a stock-out situation. We compare policy 1 with policy 7, policy 2 with policy 8, policy 3 with policy 9, policy 4 with policy 10, policy 5 with policy 11, and policy 6 with policy 12 (these policies were defined in Section 6.4). We apply ANOVA (Analysis of Variance) to test if the differences were significant. The results of these ANOVA tests can be found in Appendix G. Table 6.9 shows the cost savings that are obtained when using pipeline flexibility for those cases with cost reductions significantly different from zero. Case 2 is eliminated from the table because the stock levels at the local warehouses are equal to zero for this case. Checking the pipeline between national and local warehouses is therefore useless, since there are no normal replenishment orders. Cases 3 and 8 show no significant cost savings when checking the pipeline for outstanding orders. This is caused by the fact that the fill rates in these cases are approximately 100%, and therefore a negligible number of emergency replenishment orders is issued. The remaining cases (1, 4, 5, 6, and 7) show significant cost reductions when applying policies 1, 3, and 4. These policies have in common that they do not make use of Emergency Lateral Transhipments. Case 5 shows a significant cost reduction for all policies. This can be explained by the fact that the fill rates in case 5 are relatively low (< 80%), so many emergency shipments take place. In general we can conclude that applying the pipeline flexibility option does not decrease cost significantly when the fill rates are approximately 100%. For fill rates close to 100% pipeline flexibility provides added value when ELT flexibility is not part of the emergency replenishment structure. Finally, for low fill rates (< 80%) pipeline flexibility does provide added value for all emergency policies that were considered.

| Case | Policy | | | | | |
|------|--------|-----|------|------|------|------|
| | 1-7 | 2-8 | 3-9 | 4-10 | 5-11 | 6-12 |
| 1 | 3.8 | - | 3.7 | 3.4 | - | - |
| 3 | - | - | - | - | - | - |
| 4 | 8.7 | - | 3.5 | 3.3 | - | - |
| 5 | 13.1 | 2.3 | 3.9 | 3.6 | 2.2 | 2.0 |
| 6 | 6.6 | - | 6.5 | 6.3 | - | - |
| 7 | 16.6 | - | 14.5 | 13.9 | - | - |
| 8 | - | - | - | - | - | - |

Table 6.9: Significant cost savings (%) obtained when using pipeline flexibility

Table 6.10 shows the reduction in average response time when applying pipeline flexibility. Similar observations can be made in comparison with table 6.9. Significant response time reductions are obtained for policies 1, 3, and 4 for all cases. Savings close to 50% can be obtained; see case 7. In case 5 (with low fill rates, and therefore many emergency shipments) all policies show added value with respect to response time reduction.

| Case | Policy | | | | | |
|------|--------|-----|------|------|------|------|
| | 1-7 | 2-8 | 3-9 | 4-10 | 5-11 | 6-12 |
| 1 | 21.0 | - | 20.6 | 19.9 | - | - |
| 3 | 19.0 | - | 19.0 | 19.0 | - | - |
| 4 | 18.0 | - | 10.4 | 10.6 | - | - |
| 5 | 27.8 | 3.9 | 11.7 | 11.1 | 4.7 | 3.2 |
| 6 | 28.0 | - | 28.1 | 26.9 | - | - |
| 7 | 46.5 | - | 43.4 | 42.7 | - | - |
| 8 | 27.5 | - | 27.5 | 27.5 | - | - |

Table 6.10: Average response time reduction (%) obtained when using pipeline flexibility

With respect to the fill rate performance we can conclude from table 6.8 that applying pipeline flexibility hardly affects the fill rate performance when the fill rate values are close to 100%. However, for low fill rates (see case 5) the fill rate performance deteriorates even further with 2 to 6% when applying pipeline flexibility. This is caused by the fact that service parts that are tracked by using the pipeline option, cannot contribute to the fill rate any more: on arrival at the local warehouses these parts are allocated to customers who arrived in a stock-out situation some time ago.

6.6.2 Optimal policies

In the preceding section we investigated the added value of pipeline flexibility for all six emergency replenishment policies. In this section we identify the optimal policy (i.e. minimum-cost policy) among these policies for all eight cases under consideration. At first we restrict ourselves to policies 1 through 6, excluding pipeline flexibility. Table 6.11 shows the policies with minimum cost for the various cases. The cost savings that are obtained by applying the optimal policy are expressed as a percentage of the total cost when applying policy 1, which acts as a benchmark policy since it represents an emergency replenishment structure with a minimum level of flexibility. Table 6.11 also shows the reductions in average response time that can be obtained.

| Case | optimal policy | cost saving (%) | response time reduction (%) |
|------|----------------|-----------------|-----------------------------|
| 1 | 2,5,6 | 29.3 | 89.0 |
| 2 | 3,4,5,6 | 51.1 | 58.5 |
| 3 | 2,5,6 | 61.4 | 91.5 |
| 4 | 2,5,6 | 50.0 | 89.5 |
| 5 | 2,5,6 | 35.8 | 68.5 |
| 6 | 2,5,6 | 24.3 | 91.3 |
| 7 | 2,5,6 | 36.8 | 90.5 |
| 8 | 2,5,6 | 33.3 | 91.8 |

Table 6.11: Optimal policies with associated cost and response time reductions

We observe major cost savings and response time reductions when comparing the optimal policy with policy 1 (i.e. the benchmark policy). In six cases (1, 3, 4, 6, 7, 8) policies 2, 5, and 6 are optimal; they produce identical results. These cases are characterized by high fill rates (> 94%), so the only emergency flexibility that is applied are lateral transshipments. Direct shipments from national or central warehouses is possible (policy 5 or 6), but is not used at all. Applying emergency lateral transshipments is sufficient to deal with customers who arrive in a stock-out situation. In case 5 the optimal policies are also 2, 5, and 6, but here the simulation results are not identical. ANOVA, however, shows that the differences in performance between these policies is not significant (see Appendix G). Finally, in case 2 the optimal policies are 3, 4, 5, and 6. Note that in this case lateral transshipments are not possible, since the local stock levels are zero: policies 3 and 5 are identical, and so are policies 4 and 6. The differences between these two groups are not statistically significant (see Appendix G).

In general, we can conclude from these results that ELT-flexibility is sufficient to realize maximum cost savings. When the fill rates are high enough, other flexibilities will not be used, even when the emergency structure provides such a possibility. When the fill rates are low, other flexibilities will be used, but the added value in terms of lower cost is not statistically significant. When ELT-flexibility cannot be applied (e.g., local stock levels are zero), using direct shipments from the national warehouse realizes maximum cost savings. Also allowing direct shipments from the central warehouse does not realize significant cost savings.

So far we have restricted ourselves to policies without pipeline flexibility (policies 1 through 6). If we look at the added value of using pipeline flexibility for the optimal policies, we conclude that for these policies pipeline flexibility has added value in terms of cost savings which, however, is not statistically significant or is very small compared with the other policies. Apparently, making

Policy evaluation

use of pipeline flexibility is beneficial only when the emergency replenishment policy that is chosen, is not the optimal one!

We end this section on optimality with a remark about the effect on the fill rates when using flexibility. Flexibility (either pipeline, lateral transshipments or direct shipments) always affects the fill rate performance in a negative way. Application of any kind of flexibility means that a service part that is located somewhere in the SPSS is reserved for some customer who arrived in a stock-out situation at some local warehouse. Therefore this particular service part can no longer contribute (directly or indirectly) to the fill rate performance of any local warehouse. However, the negative effect of using flexibility on the fill rate performance is negligible when the fill rates are high (close to 100%). When the fill rates are rather low, using flexibility can deteriorate the fill rate performance significantly. This observation stresses the importance of using other performance measures (such as response time) in combination with the fill rate measure when applying flexibility.

6.6.3 Parameter main effects

We are interested in effects of the five network parameters on the performance of the different policies. Table 6.12 gives the estimated main effects (i.e. correlation coefficients) of the five network parameters on the total cost of the first six policies.

| Network parameter | Policy | | | | | |
|-------------------|--------|------|------|------|------|------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| λ | 0.43 | 0.40 | 0.42 | 0.42 | 0.41 | 0.41 |
| C | 0.65 | 0.66 | 0.61 | 0.61 | 0.64 | 0.64 |
| E | 0.16 | 0.17 | 0.17 | 0.17 | 0.16 | 0.16 |
| h | 0.42 | 0.40 | 0.41 | 0.41 | 0.41 | 0.41 |
| p | 0.30 | 0.31 | 0.34 | 0.34 | 0.32 | 0.33 |

Table 6.12: Main effects of network parameters on total cost for every policy

Two important observations can be made when reading table 6.12. First, the effect of the individual network parameters on the cost performance is practically identical for all six policies. Second, the unit price C of a service part shows the highest effect on the cost performance, whereas the emergency structure E shows the lowest effect. It is important to note however that these effects of course depend to a large extent on the extreme parameter settings, which were presented in tables 6.1 and 6.2. The results should therefore be interpreted with care (see Kleijnen and Van Groenendaal, 1992, page 178).

Table 6.12 shows the relation between the individual network parameters and the total cost for each policy. The combination of network parameters determines the stock levels at the different inventory locations in the network; see table 6.5. It is therefore also interesting to investigate the effect of these stock levels on the total cost for every policy. The results are presented in table 6.13.

| Stock levels | Policy | | | | | |
|--------------|--------|------|------|------|------|------|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| S_{cw} | 0.41 | 0.38 | 0.41 | 0.41 | 0.40 | 0.40 |
| S_{nw} | 0.15 | 0.12 | 0.15 | 0.15 | 0.14 | 0.14 |
| S_{lw} | 0.02 | 0.01 | 0.06 | 0.06 | 0.03 | 0.04 |
| System Stock | 0.14 | 0.12 | 0.17 | 0.17 | 0.14 | 0.15 |

Table 6.13: Effects of stock levels on total cost for every policy

Again we observe that the effects are practically identical for all six policies. The only significant effect on total cost is caused by the stock level at the central warehouse: a high central stock level corresponds with a high cost level for operating the SPSS. These situations correspond with a high demand rate which necessitates a high central inventory. The other stock level values (national, local, and aggregate) show an insignificant main effect on the cost behavior of the six policies.

6.6.4 The emergency replenishment structure

In the preceding experiment we compared the performance of various flexibility policies that are applied in case of a stock-out situation at a local warehouse. These policies differ from each other with respect to the number and type of flexibility options that can be applied. For all policies, however, we assume an explicit order in which these flexibility options are applied. For example, when applying policy 6, we assume that in case of a stock-out situation, we first try to apply ELT-flexibility, second direct shipment flexibility from the national warehouse, third direct shipment flexibility from the central warehouse, and fourth direct shipment flexibility from the factory. The flexibilities are applied in this order because of increasing costs and increasing response times (see table 6.2). In practice, however, it is very well possible that a direct shipment from the national (or central) warehouse is faster and cheaper than a lateral transshipment between local warehouses. This can be the case for example when daily routine replenishment shipments are made from the national (or central) warehouse to the local warehouses. In such a situation, emergency orders for local warehouses can make use of these routine shipments. In this section we investigate how this affects the performance of the system. We consider the following three scenarios, in which we apply four types of flexibility (ELT, DD-N, DD-C, DD-F) in a different order:

Policy evaluation

Scenario A: 1. ELT
2. DD-N
3. DD-C
4. DD-F

Scenario B: 1. DD-N
2. ELT
3. DD-C
4. DD-F

Scenario C: 1. DD-N
2. DD-C
3. ELT
4. DD-F

Scenario A is identical to policy 6 (see table 6.3). Scenario B prescribes that first a direct delivery from the national warehouse should be considered before using a lateral transshipment. Scenario C prescribes that a lateral transshipment should only be applied when both the national and central warehouse is out of stock. The flexibility options in each scenario are ranked in increasing cost and increasing response time. We consider the same eight cases as before, with identical parameter settings (see tables 6.1 and 6.2). Changing the order in which the various flexibility options are applied, can result in a different allocation of inventory in the system. In table 6.14 we calculated the optimal stock levels for the three scenarios, using the Muckstadt-Thomas model.

| Case | Scenario A | | | Scenario B | | | Scenario C | | |
|------|------------|----------|----------|------------|----------|----------|------------|----------|----------|
| | S_{cw} | S_{nw} | S_{lw} | S_{cw} | S_{nw} | S_{lw} | S_{cw} | S_{nw} | S_{lw} |
| 1 | 12 | 3 | 3 | 12 | 5 | 2 | 14 | 4 | 2 |
| 2 | 2 | 1 | 0 | 3 | 1 | 0 | 4 | 0 | 0 |
| 3 | 11 | 4 | 5 | 11 | 4 | 5 | 13 | 3 | 5 |
| 4 | 3 | 2 | 1 | 3 | 2 | 1 | 4 | 1 | 1 |
| 5 | 10 | 3 | 1 | 13 | 4 | 0 | 17 | 2 | 0 |
| 6 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 |
| 7 | 13 | 4 | 3 | 13 | 4 | 3 | 13 | 4 | 3 |
| 8 | 3 | 1 | 3 | 3 | 1 | 3 | 3 | 1 | 3 |

Table 6.14: Optimal stock levels for the central warehouse (S_{cw}), the national warehouses (S_{nw}), and the local warehouses (S_{lw}), according to the Muckstadt-Thomas model.

In the Muckstadt-Thomas model the use of lateral shipments is not taken into account. Scenarios B and C therefore represent situations in which the direct deliveries are faster and cheaper than in scenario A. Consequently, the optimal allocation of stock in scenarios B and C is more centralized than in scenario A. The simulation results for the eight cases with respect to total cost, average response time, and fill rate performance are displayed in table 6.15.

The optimal stock levels are identical under all scenarios for cases 6, 7, and 8 (see table 6.14). For these cases we see that the average response time increases significantly under scenarios B and C. This is caused by the fact that the inventory at the national and central warehouses is primarily used for replenishing local warehouses. Direct deliveries in case of a stock-out are often not possible, and therefore lateral shipments (longer duration and more expensive) are applied after all. Similar

observations can be made for cases 3 and 4. Note that for these cases in scenario C the optimal allocation of inventory is more centralized than in scenarios A and B. In cases 1, 2, and 5 we see that the optimal allocation of stock is different for the three scenarios. In case 1 both the average response time and the total cost increase in scenario B and C. In case 2 the average response time and total cost first decrease (scenario B) and then increase (scenario C). Since the local stock levels are zero in this case, lateral transshipments are not possible anyway. Scenario B therefore results in lower cost and faster response than scenario A. However, in scenario C the national stock levels are also equal to zero. This results in longer response times and higher penalty cost. Finally, case 5 shows a significant reduction of the average response time (i.e., approximately 50%) in scenarios B and C. This is caused by the fact that under these scenarios all stock is centralized in the direction of the national and central warehouses. Consequently, the fill rate performance in scenarios B and C is equal to zero.

| Case | Scenario A | | | Scenario B | | | Scenario C | | |
|------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | TC | RT | FR | TC | RT | FR | TC | RT | FR |
| 1 | 1944 (±14.6) | 0.22 (±0.02) | 98.8 (±0.1) | 2032 (±25.4) | 0.29 (±0.01) | 96.7 (±0.2) | 2533 (±72.4) | 0.65 (±0.02) | 96.6 (±0.2) |
| 2 | 91 (±2.3) | 0.83 (±0.01) | 0 - | 79 (±2.6) | 0.58 (±0.01) | 0 - | 93 (±0.7) | 0.62 (±0.01) | 0 - |
| 3 | 17 (±0.9) | 0.17 (±0.00) | 100.0 (±0.0) | 20 (±1.7) | 0.33 - | 100.0 (±0.0) | 29 (±5.2) | 1.00 - | 100.0 (±0.0) |
| 4 | 9 (±0.1) | 0.21 (±0.01) | 94.4 (±0.2) | 10 (±0.1) | 0.28 (±0.01) | 94.5 (±0.2) | 12 (±0.1) | 0.75 (±0.01) | 93.2 (±0.2) |
| 5 | 735 (±6.0) | 1.26 (±0.02) | 67.1 (±0.5) | 737 (±6.8) | 0.62 (±0.01) | 0 - | 651 (±4.7) | 0.62 (±0.01) | 0 - |
| 6 | 510 (±2.8) | 0.35 (±0.03) | 99.5 (±0.0) | 525 (±3.8) | 0.69 (±0.02) | 99.5 (±0.0) | 585 (±9.3) | 1.96 (±0.05) | 99.5 (±0.0) |
| 7 | 12 (±0.2) | 0.38 (±0.03) | 99.5 (±0.1) | 12 (±0.3) | 0.68 (±0.03) | 99.5 (±0.1) | 15 (±0.6) | 1.82 (±0.06) | 99.5 (±0.1) |
| 8 | 20 (±0.3) | 0.33 - | 100.0 (±0.0) | 21 (±0.6) | 0.67 (±0.0) | 100.0 (±0.0) | 24 (±1.7) | 2.00 - | 100.0 (±0.0) |

Table 6.15: Simulation results for total cost (TC), average response time (RT), and fill rate (FR) under scenarios A, B, and C.

Evaluation of three emergency structures (A, B, and C) already shows the complex interaction of various factors and their effect on the performance measures. Changing the emergency structure can result in a different allocation of stock in the system, which can affect the availability of certain flexibilities (e.g. when local stock levels are zero, lateral transshipments are not possible). The complex relation between these factors makes it difficult to formulate general guidelines.

6.7 Conclusions

In this chapter we described and analyzed the results of a simulation experiment that was conducted to evaluate different emergency replenishment policies for a Service Part Supply System. To characterize the network structure, we identified five network parameters: demand rate, purchase price, emergency shipment time and cost, holding cost, and penalty cost. An analytic model was used to calculate the minimum-cost stock levels for these different network structures. Next we used these stock levels in our simulation experiment to evaluate *different* emergency replenishment policies. In the analytical model one specific emergency replenishment policy is assumed: if demand cannot be satisfied from stock on hand, it is satisfied by direct delivery from the national warehouse, from the central warehouse, or (as the last and least desirable option) from the factory. Therefore, the analytically calculated stock levels are optimal only for *this* specific emergency replenishment policy. Applying a different emergency replenishment policy *might* result in different minimum-cost stock levels. However, since we do not have the analytic tools to calculate minimum-cost stock levels for other emergency replenishment policies, we conducted our experiment with these analytically calculated stock levels. It is important to keep this in mind, when analyzing the numerical results from the simulation experiment.

The numerical results indicate that the use of lateral transshipments in case of a stock-out situation (ELT-flexibility) is optimal. If the fill rates at the local warehouses are high (i.e., close to 100%), other flexibilities will not be used at all, even when the policy allows such flexibility. The few customers who arrive in stock-out situations, can all be dealt with by applying lateral transshipments from other local warehouses. If the fill rates are rather low (i.e., lower than 80%), other flexibilities will be used every now and then, but their contribution in terms of cost reduction is not statistically significant. If ELT-flexibility is not possible (e.g., local stock levels are zero or geographical circumstances prevent the use of lateral transshipments), the numerical results indicate that the use of direct shipments from the national warehouse are optimal. Using additional direct shipments from the central warehouse does not contribute significantly to cost reduction.

Pipeline flexibility means checking the pipeline for outstanding normal replenishment orders that are due to arrive earlier than any other emergency procedure could realize. Its use proves to be interesting, when lateral transshipments are not allowed. In other words, if the optimal policy (allowing for ELT-flexibility) is not applied, pipeline flexibility can be very rewarding. However, when ELT-flexibility is allowed, the use of pipeline flexibility has a negligible effect on the total cost and on the average response time.

Investigation of the (main) effects of the network parameters on total cost showed that the purchase price of a service part has the highest effect (0.64), and the emergency structure (in terms of cost

and lead time) the lowest effect (0.16). The effects of stock levels on total cost showed that only the *central* stock level had a significant effect (0.40).

The emergency replenishment structure influences the performance of the system as well. Changing the order in which flexibility options are called upon, can result in a different allocation of inventory in the system (e.g. centralization of stock) and can affect the average response time for backordered customers significantly.

In this simulation experiment we concentrated on three performance measures: total cost, average response time, and fill rate. When analyzing the simulation data, we identified optimal policies as the policies with the lowest total cost. In practically all cases this cost optimizing (or minimizing) point of view corresponds with optimizing (or minimizing) the average response time. When the goal, however, is to maximize the average fill rate at the local warehouse, then a different picture emerges. From a fill rate optimizing (or maximizing) point of view, flexibility should never be implemented in the SPSS! In that situation the policy with the lowest level of flexibility (policy 1) is always optimal. The use of flexibility implies that a service part that is directly or indirectly contributing to some local fill rate, is removed from the system to serve a customer who arrived in a stock-out situation anyway. It is important to see that the optimization of *different* performance measures can result in different and *conflicting* recommendations.

Chapter 7

Conclusions and recommendations for further research

7.1 Introduction

In this thesis we investigated the use of flexibility in a repair and distribution environment for service parts. These environments are typically characterized by high value items with a low demand frequency. In case of a demand, short response times are often very crucial to the customer. Examples of such situations are grounded aircraft or shut down oil platforms at sea, as a result of missing service parts. But also in other industries, the emphasis on after sales service is increasing. High service performance towards the customer becomes more and more a competitive weapon nowadays. In practice, every company applies flexibility in case of an emergency situation. When a customer needs a specific service part and the inventory location located nearest to the customer is out of stock, alternative solutions are sought. Depending on the consequences for the customer, i.e. complete stoppage of equipment or just reduced functionality, different flexibility options are applied at different cost. Application of flexibility in emergency situations is often based on ad hoc decisions. In this thesis we tried to embed the use of flexibility in the context of the 'normal' operational control activities. We view the use of flexibility as an integral part of these operational processes.

The strategic policy of a company is a dominant factor in determining the optimal flexibility policy. If the emphasis is on cost reduction, centralization of service parts inventories leads to a reduction of inventory holding cost, replenishment cost, and obsolescence risk. However, due to an increasing distance between the customers and the (centralized) service part inventory, the service performance decreases. The flexibility policy in such a situation is focused on a fast and reliable distribution of service parts to customers spread over a large geographical area. If the emphasis is on service performance, decentralization of service parts inventories leads to a fast response to customers in case of a demand. However, inventory holding cost, replenishment cost, and obsolescence risk increase. The flexibility policy in such a situation is focused on tracking and tracing service parts in the distribution system as a whole, such that a customer demand in a stock-out situation can be satisfied as soon as possible.

The use of flexibility is very often situation dependent. Market conditions may prescribe a minimum response time in case of a breakdown at a customer site. High, contractually agreed, penalty cost may require decentralized inventories of service parts with additional high risks. Specific

circumstances, such as customs regulations or geographical conditions, may exclude the use of certain flexibility options. As a result, it is very difficult to formulate general guidelines with respect to the use of flexibility in a Service Part Supply System. The aim of this thesis is to make a contribution with respect to the applicability of flexibility in such systems. For this aim we developed a general framework (see Chapter 3) and some specific quantitative models (see Chapters 4, 5, and 6). The main conclusions are summarized in Section 7.2. In Section 7.3 we give some recommendations for further research.

7.2 Main conclusions

In this thesis we presented and discussed several models related to the design and control of Service Part Distribution Systems. We now present an overview of the main results with respect to the SPSS-framework (Chapter 3), the Emergency Repair Model (Chapter 4), the Emergency Supply Model (Chapter 5), and the simulation experiment (Chapter 6):

SPSS-framework

- 1) Tool for evaluating current logistic designs of supply systems for service parts.
- 2) Identification of flexibility options that can improve the performance.
- 3) Analysis of economic trade-offs that are associated to various flexibility options.

Emergency Repair Model

- 1) Exact analysis of emergency repair flexibility.
- 2) Modeling both the initial stock level and the trigger level as a decision variable.
- 3) Major cost reductions in comparison with other policies.
- 4) Insensitivity of service performance with respect to repair time distributions.
- 5) Sensitivity of service performance with respect to demand distribution.

Emergency Supply Model

- 1) Approximate model with good performance.
- 2) Major cost reductions in comparison with models that do not incorporate flexibility.
- 3) Insensitivity of service performance with respect to lead time distribution.
- 4) Effect of different pooling structures.

Simulation experiment

- 1) Comparison of several emergency replenishment policies with respect to cost and service.
- 2) Influence of system parameters on cost and service performance.

Conclusions

In Chapter 3 we introduced the Service Part Supply System: a framework that can be used to identify opportunities to implement flexibility with respect to the control of the repair and distribution processes of service parts. We defined three types of repair flexibility (return flow, work order release, and repair shop flexibility) and three types of supply flexibility (allocation, pooling, and direct shipment flexibility). This framework is an excellent tool for evaluating the current logistic design of practical repair and supply systems. Furthermore, it can be used to identify opportunities to improve the performance of such systems. Implementation of any kind of flexibility requires a careful economic trade-off between the pro's (in terms of increased service performance) and con's (in terms of incurred cost) of such a decision. For that reason we constructed two analytic models that can be used to support decisions with respect to the implementation of flexibility.

In Chapter 4 we introduced the Emergency Repair Model: an analytic model that can be used to investigate the economic trade-off when considering the use of emergency repair. It models a single-echelon inventory location where failed service parts, arriving according to a Poisson process, are sent either into normal repair or emergency repair. This decision is dependent on the repair trigger level: if the net inventory exceeds the trigger level, the part is sent into normal repair. Otherwise, the part is sent into (faster but more expensive) emergency repair. For a given cost structure and emergency repair speed, we can determine the optimal stock level and trigger level that minimize total cost. The numerical results show that, dependent on situational factors, positive, negative, or zero trigger levels can be cost optimal. We compared the performance of the Emergency Repair Model with the performance of three other policies: 1) no emergency repair is applied, 2) the trigger level for emergency repair is equal to zero, and 3) customers that arrive in a stock-out situation always wait for their own part that has been sent into emergency repair. Major cost savings can be obtained when comparing the performance of our Emergency Repair Model with the performance of these three alternatives. Considering both the stock level and the trigger level as a decision variable can lead to major cost reductions.

An important result that emerged from the numerical experiments with the Emergency Repair Model is the insensitivity of the performance of the model with respect to the choice of lead time distribution. Especially the fill rate measure showed a high level of insensitivity. This phenomenon can be partly explained by the fact that the repair processes, both normal and emergency repair, are modeled as independent processes. Another explanatory factor is the relatively long inter-arrival times of failed parts in comparison to the length of the repair lead times. On the other hand, the sensitivity of the numerical results with respect to the assumption of exponentially distributed inter-arrival times of failed parts, proved to very crucial. This is an important observation, since many repair models for service parts assume Poisson demand processes. This means that applying such models in situations with less erratic demand processes, e.g. in case of preventive maintenance situations, can lead to a serious underestimation of the performance.

In Chapter 5 we introduced the Emergency Supply Model: an analytic model that can be used to investigate the trade-off when considering the use of pooling flexibility and direct shipment flexibility. It models a two-echelon inventory system where, in case of a stock-out at a local warehouse, lateral transshipments from any other local warehouse are applied. If this is not possible, direct shipments from the central warehouse are applied. Finally, if all these flexibility options are not possible, direct deliveries from the factory are applied. The model is approximative in nature, but simulation results are presented that show a good performance of the approximation. Since the modeling technique is based on Markov analysis, normal replenishment lead times are assumed to be exponentially distributed. However, again we observe a high degree of insensitivity of the performance with respect to this assumption. Simulation with deterministic lead times produces almost identical results. Furthermore, we compare the cost performance of the Emergency Supply Model with the cost performance of the VARI-METRIC model, in which no flexibility is applied at all. The numerical results indicate that major cost reductions can be obtained when using supply flexibility.

We also investigated the influence of the pooling structure on the performance of the supply system. In the Emergency Supply Model our modeling technique necessitates the assumption of complete pooling, i.e., all local warehouses are potential sources for issuing a lateral transshipment in case of a stock-out situation at any local warehouse. In a simulation experiment we compared the performance of the complete pooling concept with two other pooling concepts: 1) fixed pooling (assuming that all local warehouses are divided into separate groups, a local warehouse can only make a request for a lateral transshipment to local warehouses in its own pooling group) and 2) variable pooling (assuming that all local warehouses are located on a circle around the central warehouse, a local warehouse can only make a request for a lateral transshipment to its direct neighboring local warehouses). The results indicate that for high fill rates, i.e. the stock levels at the local warehouses are high enough to guarantee a fill rate of 70% or more, the differences between these three pooling concepts are very small. For fill rates lower than 70% the differences are significant. Application of the complete pooling concept results in the highest fraction of demand at a local warehouse that is satisfied through lateral transshipments. This is logical, since all other local warehouses are candidates for sourcing such a lateral transshipment. However, the fill rate performance under the complete pooling concept is significantly worse than under the other two concepts. A local warehouse that often acts as a source for lateral transshipments, which is the case under the complete pooling concept, reduces the availability of stock for its own customers. The difference in performance between the fixed and variable pooling group, even for fill rates lower than 70%, is still negligible. We can conclude that the pooling structure is only of interest when the fraction of demand that is satisfied through lateral transshipments is very high, i.e. more than 30%. In such a situation, application of pooling flexibility can no longer be viewed as an emergency option for stock-out situations. The use of lateral transshipments becomes a routine activity,

Conclusions

comparable to the normal replenishment shipments from the central warehouse.

In Chapter 6 we compared the cost and service performance of various flexibility policies. The analytic models presented in Chapters 4 and 5 represent two specific flexibility policies. In the Emergency Repair Model we prescribe that an emergency repair is issued if the net stock at the inventory location is lower than a specific trigger level. Furthermore, it is assumed that a waiting customer receives the first part that becomes available from repair, either normal or emergency. In the Emergency Supply Model we prescribe that, in case of a stock-out situation at a local warehouse, first the possibility of using lateral transshipments is checked, second the possibility of using direct deliveries from the central warehouse is checked, and third, the possibility of using direct deliveries from the factory is checked (which is always possible). These specific flexibility policies in the Emergency Repair Model and Emergency Supply Model are necessary from a modeling point of view. In Chapter 6 we used simulation to evaluate more complex flexibility policies in a three-echelon network. We considered twelve flexibility policies that contain one or more of the following flexibility options in case of a stock-out situation at a local warehouse: pipeline flexibility (waiting for replenishment order that is due to arrive), pooling flexibility (applying lateral transshipments between local warehouses), and direct shipment flexibility (applying direct deliveries from the national warehouse, the central warehouse, or the factory).

The results of the simulation experiment should be interpreted very carefully. The reason for this is that we used an analytic model (i.e., the Muckstadt-Thomas model) to derive 'optimal' stock levels that were implemented in the simulation model. Since the Muckstadt-Thomas model does not include the possibility of lateral transshipments, the resulting stock levels are not necessarily optimal for the various flexibility policies in the simulation experiment. More precisely, the optimal stock allocation that results from the Muckstadt-Thomas model is expected to be more centralized, since flexibility at the lowest level (i.e. local warehouse level) is not possible. The main goal of the experiment is merely to illustrate the complexity of flexibility strategies and the effects of important parameters.

From a cost point of view, application of pooling flexibility is always optimal. If the fill rates are very high (i.e., close to 100%), other flexibilities that are present will not be used. If the fill rates are rather low (i.e., lower than 80%), other flexibilities will be used every now and then, but the effect on the cost performance is negligible. If pooling flexibility is not part of the flexibility policy, application of direct shipment flexibility from the national warehouse is optimal. Additional direct deliveries from the central warehouse does not affect the performance significantly. The use of pipeline flexibility, i.e. waiting for a replenishment order that is due to arrive instead of issuing some kind of emergency order, is only worth considering when pooling flexibility is not part of the flexibility policy. If pooling flexibility is an option, which is optimal, pipeline flexibility does not

contribute significantly to the performance of the system. We also investigated the effect of the emergency structure on the performance of the supply system. Changing the order in which different flexibility options are applied can have a significant effect on the service performance, especially the average length of the response time.

The different models that have been presented in this thesis can all be used to make an economic trade-off when considering the use of repair or supply flexibility in a Service Part Supply System. The advantage of applying the analytic models is that they can be used to find, within a reasonable time, optimal values for decision variables (e.g. stock levels, trigger levels, emergency lead times). However, these models are restrictive in nature, since they assume a specific flexibility policy in case of an emergency. That is, the analytic models are not very flexible. The simulation model that has been presented, can be adapted in such a way that it takes into account many situational factors. It is therefore more flexible than analytic models. However, the disadvantage of the simulation model is that it is difficult to use as an optimization tool. It can merely be used as an instrument to investigate the effects of relevant parameters. A sound balance between the use of analytic models and simulation models is therefore necessary to investigate the control and design issues of service part distribution systems.

Finally we want to stress the strong relation between design and control issues of Service Part Distribution Systems. Designing a supply network for service parts without considering the operational control policies can result in unnecessarily high costs. It is therefore very important to consider design and control of these systems simultaneously.

7.3 Topics for further research

The analytic models presented in this thesis provide some interesting topics for further research. In the Emergency Repair Model we assume that repair times are independent of each other by modeling the repair processes as infinite server queues. We already showed how the normal repair process can be modeled as an $M/M/1$ queue (Section 4.4.5) with limited repair capacity. Analyzing the effect of limited capacity in such a system, especially the normal repair process, provides an interesting topic of research. Furthermore, the analysis can be extended to general $M/M/c$ queues for the normal repair process. Extending the analysis of the Emergency Repair Model to a multi-echelon situation is also very interesting. This represents a situation in which there are several local warehouses that send failed service parts to a central repair facility for normal repair. When the net inventory at a local warehouse is lower than a specific trigger level, the failed part is sent to a local vendor for emergency repair.

Conclusions

Extending the analysis of the Emergency Supply Model to three or more echelon networks is also an interesting topic of further research. This is not easy, since we have to assume complete pooling in our analysis. This means that analysis of a three-echelon system requires a multi-dimensional Markov state space with state variables for the inventory positions at every pooling group, at every national warehouse, and at the central warehouse. Next to the dimensionality problem, transitions between these state variables become very complex. Another research issue that can be investigated with respect to the Emergency Supply Model, is the pooling structure. We assume that a lateral transshipment is issued from a randomly chosen local warehouse with positive stock on hand. It is possible to model other pooling structures as well, in which some kind of priority list is used with respect to the local warehouses that may act as source for lateral shipments. Any kind of priority list may be applied, as long as the complete pooling concept is not violated: a direct delivery from the central warehouse is only made when all local warehouses are out of stock.

The use of the simulation model illustrated the complex relations between different system parameters and the effects on the cost and service performance. For a detailed analysis of various flexibility policies and the determination of optimal solutions, analytical models are needed that incorporate these various flexibility options. Especially the sensitivity of the performance (both cost and service) of various policies on the individual system parameter needs to be investigated in more detail. Our simulation experiment was primarily focussed on comparing the performance of various policies and not on the influence of individual parameters on the performance. Future research is needed to determine the impact of individual system parameters on cost and service performance of various flexibility policies.

References

- ABBOUD, N.E. (1996), The Markovian Two-echelon Repairable Item Provisioning Problem. *Journal of the Operational Research Society* 47, 284-296.
- AGGARWAL, P.K., and K. MOINZADEH (1994), Order Expedition in Multi-echelon Production/Distribution Systems. *IIE Transactions* 26 (2), 86-96.
- AHMED, M.A., D. GROSS, and D.R. MILLER (1992), Control Variate Models for Estimating Transient Performance Measures in Repairable Item Systems. *Management Science* 38 (3), 388-399.
- ALBRIGHT, S.C., and A. SONI (1988), Markovian Multi-echelon Repairable Inventory System. *Naval Research Logistics* 35, 49-61.
- ALFREDSSON, P. (1996), Optimization of a Multi-echelon Repairable Item Inventory System with Simultaneous Repair Facility Localization. To appear in the *European Journal of Operational Research*.
- ALFREDSSON, P., and J.H.C.M. VERRIJDT (1996), Modeling Emergency Supply Flexibility in a Two-echelon System. Research report TUE/TM/LBS/96-01, Eindhoven University of Technology (submitted for publication).
- AXSÄTER, S. (1990), Modelling Emergency Lateral Transshipments in Inventory Systems. *Management Science* 36 (11), 1329-1338.
- BARANKIN, E.W. (1961), A Delivery-lag Inventory Model with an Emergency Provision. *Naval Research Logistics Quarterly* 8, 285-311.
- BEMELMANS, T.M.A. (1987), *Bestuurlijke informatiesystemen en automatisering*. Stenfert Kroese, Leiden, the Netherlands (in Dutch).
- BERTRAND, J.W.M., J.C. WORTMANN, and J. WIJNGAARD (1990), *Production Control: A Structural and Design Oriented Approach*. Elsevier Science Publishers B.V., Amsterdam.
- BÜYÜKKURT, M.D., and M. PARLAR (1993), A Comparison of Allocation Policies in a Two-echelon Repairable-item Inventory Model. *International Journal of Production Economics* 29, 291-302.
- CHEUNG, K.L., and W.H. HAUSMAN (1995), Multiple Failures in a Multi-item Spares Inventory Model. *IIE Transactions* 27, 171-180.
- CHO, D.I., and M. PARLAR (1991), A Survey of Maintenance Models for Multi-unit Systems. *European Journal of Operational Research* 51, 1-23.
- CHUA, R.C.H., G.D. SCUDDER, and A.V. HILL (1993), Batching Policies for a Repair Shop with Limited Spares and Finite Capacity. *European Journal of Operational Research* 66, 135-147.
- CLARK, A.J. (1981), Experiences with a Multi-indentured, Multi-echelon Inventory Model. In: *Multi-level Production/Inventory Control Systems: Theory and Practice* (ed. L.B. Schwarz), North-Holland Publishing Company, 299-330.

- COHEN, M.A., P.R. KLEINDORFER, and H.L. LEE (1986), Optimal Stocking Policies for Low Usage Items in Multi-echelon Inventory Systems. *Naval Research Logistics Quarterly* 33, 17-38.
- COHEN, M.A., P.R. KLEINDORFER, and H.L. LEE (1988), Service Constrained (s,S) Inventory Systems with Priority Demand Classes and Lost Sales. *Management Science* 34, 482-499.
- COHEN, M.A., P.R. KLEINDORFER, and H.L. LEE (1989), Near-optimal Service Constrained Stocking Policies for Spare Parts. *Operations Research* 37, 104-117.
- COHEN, M.A., and H.L. LEE (1990), Out of Touch with Customer Needs? Spare Parts and After Sales Service. *Sloan Management Review*, Winter 1990, 55-66.
- COHEN, M., P.V. KAMESAN, P. KLEINDORFER, H. LEE, and A. TEKERIAN (1990), Optimizer: IBM's Multi-echelon Inventory System for Managing Service Logistics. *Interfaces* 20, 65-82.
- COHEN, M.A., P.R. KLEINDORFER, H.L. LEE, and D.F. PYKE (1992), Multi-item Service Constrained (s,S) Policies for Spare Parts Logistics Systems. *Naval Research Logistics* 39, 561-577.
- COOPERS & LYBRAND and AFISM INTERNATIONAL (1991), Service Operations Strategies of the 90s. Research report.
- CORBET, M.H. (1995), *Logistiek Management & Management Accounting*. Ph.D. Dissertation, Eindhoven University of Technology (in Dutch).
- DADA, M. (1992), A Two-echelon Inventory System with Priority Shipments. *Management Science* 38, 1140-1153.
- DANIEL, K.H. (1962), A Delivery-lag Inventory Model with Emergency Order. *Multistage Inventory Models and Techniques*, Chapter 2, Scarf, Gilford, Shelly (eds.), Stanford University Press, Stanford, Ca.
- DARYANANI, S., and D.R. MILLER (1992), Calculation of Steady-state Probabilities for Repair Facilities with Multiple Sources and Dynamic Return Priorities. *Operations Research* 40 (2), S248-S256.
- DEKKER, R., M.J. KLEIJN, and P.J. DE ROOY (1996), A Spare Parts Stocking Policy Based on Equipment Criticality. Research report 9625/A. Econometric Institute, Erasmus University Rotterdam, the Netherlands.
- EBELING, C.E. (1991), Optimal Stock Levels and Service Channel Allocations in a Multi-item Repairable Asset Inventory System. *IIE Transactions* 23 (2), 115-120.
- ERNST, R., and M.A. COHEN (1993), Dealer Inventory Management Systems. *IIE Transactions* 25, 36-49.
- FEENEY, G.J., and C.C. SHERBROOKE (1966), The (S-1,S) Inventory Policy under Compound Poisson Demand. *Management Science* 12 (5), 391-411.

References

- FORTUIN, L. (1980), The All-time Requirement of Spare Parts for Service After Sales - Theoretical Analysis and Practical Results. *International Journal of Operations and Production Management* 1 (1), 59-70.
- FORTUIN, L. (1981), Reduction of the All-time Requirement for Spare Parts. *International Journal of Operations and Production Management* 2 (1), 29-37.
- FORTUIN, L. (1984), Initial Supply and Re-order Level of New Service Parts. *European Journal of Operational Research* 15, 310-319.
- FUKUDA, Y. (1964), Optimal Policy for the Inventory Problem with Negotiable Leadtime. *Management Science* 10, 690-708.
- GITS, C.W. (1984), *On the Maintenance Concepts for a Technical System: A Framework for Design*. Ph.D. Dissertation, Eindhoven University of Technology.
- GRAVES, S.C. (1985), A Multi-echelon Inventory Model for a Repairable Item with One-for-one Replenishment. *Management Science* 31 (10), 1247-1256.
- GROSS, D., R.M. SOLAND, and C.E. PINKUS (1981), Designing a Multi-product, Multi-echelon Inventory System. In: *Multi-level Production/Inventory Control Systems: Theory and Practice* (ed. L.B. Schwarz), North-Holland Publishing Company, 11-49.
- GROSS, D. (1982), On the Ample Service Assumption of Palm's Theorem in Inventory Modeling. *Management Science* 28 (9), 1065-1079.
- GROSS, D., D.R. MILLER, and R.M. SOLAND (1983), A Closed Network Model for Multi-echelon Repairable Item Provisioning. *IIE Transactions* 15 (4), 344-352.
- GROSS, D., and D.R. MILLER (1984), Multi-echelon Repairable-item Provisioning in a Time-varying Environment Using the Randomization Technique. *Naval Research Logistics Quarterly* 31, 347-361.
- HAAS, H.F.M. DE (1995), The Coordination of Initial Stock and Flexible Manpower in Repairable Item Systems. *Ph.D. Thesis*, Eindhoven University of Technology, the Netherlands.
- HAAS, H.F.M. DE, and J.H.C.M. VERRIJDT (1996), Target Setting for the Departments in an Aircraft Repairable Item System. To appear in the *European Journal of Operational Research*.
- HAUSMAN, W.H., and G.D. SCUDDER (1982), Priority Scheduling Rules for Repairable Inventory Systems. *Management Science* 28 (11), 1215-1232.
- HAUSMAN, W.H., and N.K. ERKIP (1994), Multi-echelon vs. Single-echelon Inventory Control Policies for Low-demand Items. *Management Science* 40 (5), 597-602.
- HEYMAN, G.J.L. (1996), *Interne logistieke beheersing van spare parts bij Philip Morris Holland B.V. Master Thesis no. 2867*, Faculty of Technology Management, Eindhoven University of Technology (in Dutch).

- HULL, D.L., and J.F. COX (1994), The Field Service Function in the Electronics Industry: Providing a Link between Customers and Production/Marketing. *International Journal of Production Economics* 37, 115-126.
- KANTERS, M.M.M. (1994), To Spare or not to Spare? That's the (key) question! *Master Thesis* no. 2601, Faculty of Technology Management, Eindhoven University of Technology.
- KAPLAN, A., and D. ORR (1985), An Optimum Multi-echelon Repair Policy and Stockage Model. *Naval Research Logistics Quarterly* 32, 551-566.
- KLEIJNEN, J.P.C., and W. VAN GROENENDAAL (1992), *Simulation: A Statistical Perspective*. John Wiley & sons.
- LAGODIMOS, A.G., A.G. DE KOK, and J.H.C.M. VERRIJDT (1995), The Robustness of Multi-echelon Service Models under Autocorrelated Demands. *Journal of the Operational Research Society* 46 (1), 92-103.
- LAMERS, J. (1994), Field Service Spare Part Logistics at Intergraph. *Master Thesis* no. 2443, Faculty of Technology Management, Eindhoven University of Technology.
- LAW, A.M., and W.D. KELTON (1991), *Simulation Modeling and Analysis*. McGraw - Hill.
- LEE, H.L. (1987), A Multi-echelon Inventory Model for Repairable Items with Emergency Lateral Transshipments. *Management Science* 33, 1302-1316.
- LEVITT, T. (1983), After the Sale is Over. *Harvard Business Review*, September-October 1983, 87-93.
- MANN JR., L. (1983), *Maintenance Management*, Lexington Books, MacMillan, Inc.
- 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.
- MIRASOL, N.M.(1964), A Systems Approach to Logistics. *Operations Research* 12, 707-724.
- MOINZADEH, K., and H.L. LEE (1986), Batch Size and Stocking Levels in Multi-echelon Repairable Systems. *Management Science* 32 (12), 1567-1581.
- MOINZADEH, K., and S. NAHMIAS (1988), A Continuous Review Model for an Inventory System with Two Supply Modes. *Management Science* 34, 761-773.
- MOINZADEH, K., and C.P. SCHMIDT (1991), An (S-1,S) Inventory System with Emergency Orders. *Operations Research* 39, 308-321.
- MUCKSTADT, J.A. (1973), A Model for Multi-item, Multi-echelon, Multi-indenture Inventory System. *Management Science* 20, 472-481.
- MUCKSTADT, J.A. (1978), Some Approximations in Multi-item, Multi-echelon Inventory Systems for Recoverable Items. *Naval Research Logistics Quarterly* 25, 377-394.
- MUCKSTADT, J.A. (1979), A Three-echelon, Multi-item Model for Recoverable Items. *Naval Research Logistics Quarterly* 26, 199-221.
- MUCKSTADT, J.A., and L.J. THOMAS (1980), Are Multi-echelon Inventory Methods Worth Implementing in Systems with Low-demand-rate Items? *Management Science* 26, 483-494.

References

- NAHMIAS, S. (1981), Managing Repairable Item Inventory Systems: A Review. In: L.B. Schwarz (ed.), *Multi-level Production/Inventory Control Systems: Theory and Practice*, TIMS Studies in the Management Sciences, 16, Elsevier, 253-278.
- NEUTS, M.F. (1981), *Matrix-geometric solutions in stochastic models*, John Hopkins University Press, Baltimore.
- PALM, C. (1938), Analysis of the Erlang Traffic Formula for Busy-signal Arrangements. *Ericsson Technics* 5, 39-58.
- PRINS, E.J.A. (1996), Distribution Strategies in European Spare Part Logistics. *Master Thesis* no. 2999, Faculty of Technology Management, Eindhoven University of Technology.
- PYKE, D.F. (1990), Priority Repair and Dispatch Policies for Repairable-item Logistics Systems. *Naval Research Logistics* 37, 1-30.
- RAMASWAMI, V., and P.G. TAYLOR (1996), Some Properties of the Rate Operators in Level Dependent Quasi-birth-and-death Processes with a Countable Number of Phases. *Stochastic Models* 12, 143-164.
- ROSENSHINE, M., and D. OBEE (1976), Analysis of a Standing Order Inventory System with Emergency Orders. *Operations Research* 24, 1143-1155.
- SCHNEEWEISS, C.A., and H. SCHRÖDER (1992), Planning and Scheduling the Repair Shops of the Deutsche Lufthansa AG: A Hierarchical Approach. *Production and Operations Management* 1, 22-33.
- SCUDDER, G.D., and W.H. HAUSMAN (1982), Spares Stocking Policies for Repairable Items with Dependent Repair Times. *Naval Research Logistics Quarterly* 29, 303-322.
- SCUDDER, G.D. (1984), Priority Scheduling and Spares Stocking policies for a repair shop: the multiple failure case. *Management Science* 30, 739-749.
- SENGERS, E. (1995), Het Optimaliseren van de Voorraad Spare Parts bij Siemens. *Master Thesis* no. 2773, Faculty of Technology Management, Eindhoven University of Technology (in Dutch).
- SHANKER, K. (1981), Exact Analysis of a Two-Echelon Inventory System for Recoverable Items under Batch Inspection Policy. *Naval Research Logistics Quarterly* 28, 579-601.
- SHERBROOKE, C.C. (1968), METRIC: A Multi-echelon Technique for Recoverable Item Control. *Operations Research* 16, 122-141.
- SHERBROOKE, C.C. (1971), An Evaluator for the Number of Operationally Ready Aircraft in a Multi-echelon Availability Model. *Operations Research* 19, 618-635.
- SHERBROOKE, C.C. (1986), VARI-METRIC: Improved Approximations for Multi-indenture, Multi-echelon Availability Models. *Operations Research* 34, 311-319.
- SHERBROOKE, C.C. (1992a), Multi-echelon inventory systems with lateral supply. *Naval Research Logistics* 39, 29-40.
- SHERBROOKE, C.C. (1992b), *Optimal inventory modelling of systems: multi-echelon techniques*. John Wiley & sons inc., New York.

- SIMON, R.M. (1971), Stationary Properties of a Two-echelon Inventory Model for Low Demand Items. *Operations Research* 19, 761-773.
- SLAY, F.M. (1984), VARI-METRIC: An Approach to Modelling Multi-echelon Resupply When the Demand Process is Poisson with a Gamma Prior. Logistics Management Institute, Washington D.C., Report AF301-3.
- SMITH, M.A.J., and R. DEKKER (1996), On the (S-1,S) Model for Renewal Demand Processes: Poisson's Poison. Research report 9608/A, Econometric Institute, Erasmus University Rotterdam (submitted for publication).
- STAHL, M.J., and D.W. GRIGSBY (1992), *Strategic Management for Decision Making*. PWS-Kent Publishing Company, Boston.
- TAKÁCS, L. (1962), *Introduction to the theory of queues*. Oxford University Press, New York.
- TEUNTER, R.H., and W.K. KLEIN HANEVELD (1995), Optimal Provisioning Strategies for Slow Moving Spare Parts with Small Lead Times. To appear in *Journal of the Operational Research Society*.
- TEUNTER, R.H., and L. FORTUIN (1996), End-of-Life Service. Research report, Graduate School/Research Institute Systems Organization and Management, University of Groningen, the Netherlands (submitted for publication).
- TIJMS, H.C. (1988), *Stochastic Modelling and Analysis: A Computational Approach*. John Wiley & sons inc., New York.
- VERRIJDT, J.H.C.M. (1995), Flexibility Trade Off in a Service Part Supply System. *Proceedings of the seventh NOBO research day in Groningen*, 183-189.
- VERRIJDT, J.H.C.M., and A.G. DE KOK (1995), Distribution Planning for a Divergent N-echelon Network without Intermediate Stocks under Service Restrictions. *International Journal of Production Economics* 38, 225-243.
- VERRIJDT, J.H.C.M., and A.G. DE KOK (1996), Distribution Planning for a Divergent Depotless Two-echelon Network under Service Constraints. *European Journal of Operational Research* 89, 341-354.
- VERRIJDT, J.H.C.M., I. ADAN, and A.G. DE KOK (1996), A Trade-off between Emergency Repair and Inventory Investment. Research report TUE/TM/LBS/95-05, Eindhoven University of Technology (submitted for publication).
- WHITTMORE, A.S., and S. SAUNDERS (1977), Optimal Inventory under Stochastic Demand with Two Supply Options. *SIAM Journal for Applied Mathematics* 32, 293-305.
- WOLFF, R.W. (1982), Poisson arrivals see time averages. *Operations Research* 30, 223-231.
- WOUTERS, M.J.F. (1993), *Relevant Costs or Full Costs? Explaining Why Managers Use Capacity Cost Allocations for Short-Term Decisions*. Ph.D. Dissertation, Eindhoven University of Technology.

Appendix A

The Parts Business Forum

In the period this research was conducted, the author has been a member of the Parts Business Forum (PBF). This is an industrial forum in which managers from a wide range of companies, academics, and consultants meet on a regular basis to discuss issues of present interest in the field of service part management. Although the nature of these issues is broader than the topics addressed in this book, attending these meetings proved to be a fruitful and interesting frame of practical reference. In this appendix we elaborate on this industrial forum and its objectives. We also give an overview of the topics that were discussed in these meetings and the general findings.

The first meetings of the Parts Business Forum were held in 1993. It is a cooperative effort of several organizations with the objective to improve the performance of the service part business of industrial companies. Among its first members are Districon, a consultancy firm that initiated these meetings, the faculty of Technology Management of the Eindhoven University of Technology, and seven industrial companies (i.e., Daf Trucks, Fokker Aircraft, Greenland, Honeywell, Louwman & Parqui, Océ, and Siemens). At present the number of participants is approximately twenty. The strategic aim of the Parts Business Forum is to realize an efficient and effective management of the parts business, such that customer satisfaction is maximized and costs and capital investments are minimized. The means to realize this goal exist of presentations by experts, company visits, scientific research, and workshops.

The initial members of the PBF composed a document in which the main topics of interest were identified: parts supply in relation to corporate goals, supply chain effects, operational control, and information technology. Furthermore, a number of desirable outputs were formulated, such as improved tuning between production logistics and service part logistics, supply chain control of service parts, reduction of response times to service organizations, and exchangeability of information. In order to compare the service part businesses of the various companies and to establish a general frame of reference, the PCOI-model (Bemelmans, 1987) was used to characterize the individual companies. According to this model, first the primary Processes (P) have to be identified, next the Control (C) of these processes, and finally the Organization (O) and Information (I) aspects. Next to these four aspects, the product-market combination (the initial product for which the service parts are needed and its market situation) was identified as discerning factor. This approach also served as an inspiration for the logo of the Parts Business Forum (see figure A.1).

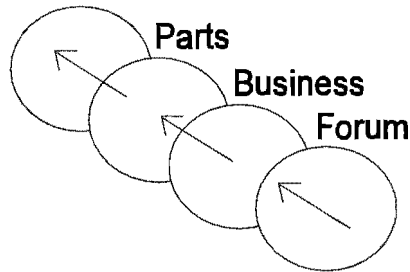


Figure A.1: The Parts Business Forum logo

Several issues related to service part management have been topic of discussion in recent years. Here we we give a brief overview of these issues and the general outcomes of the discussions. It is not our intention to give a detailed report of these meetings, but merely to present some important topics in service part management nowadays.

Strategic developments

In this meeting two important strategic issues were discussed: customer service and market penetration. In practice companies offer two types of products: the hardware (i.e., the initial product such as an airplane or a truck) and the software (i.e., customer service as added value for the initial product). Customer service contains several aspects, such as documentation and training facilities. The customer service performance can be measured objectively (e.g. in terms of fill rates or response times) and subjectively (e.g. interviewing the customer). Consolidating and increasing market share is essential. The revenues from the parts business can be interpreted as the product of volume and net margin. The net margin is low and fixed, due to high competitive pressure. The volume is decreasing as a result of increasing product reliability. Increasing the assortment is a way to maintain the total revenues. Offering parts of lower quality, the so-called second line parts, is another way. The latter option, however, is in contradiction with the general trend of quality improvement.

Customer satisfaction

The importance of customer satisfaction has already been mentioned. One of the PBF members explained the method that was used to measure customer satisfaction. The aim is to increase the satisfaction from a customer's perspective and hence influence his or her purchase behavior. This is done for both 'Sales' (responsible for the initial product) and 'After Sales' (responsible for the service parts). The timing of executing the satisfaction measurement is very important. 'Sales'

Appendix

measures the customer satisfaction immediately after the purchase has taken place. 'After Sales' on the other hand, measures the customer satisfaction annually for a period of five years. Based on these findings, the dealers (who are in direct contact with the customers) are judged and an action list is composed to improve the weak points. Next to this customer satisfaction research within one company, a similar research is conducted at branch level.

Life cycle issues

In Chapter 1 we already discussed the various phases of the product life cycle and the associated service part life cycle. In a workshop some general guidelines were formulated with respect to the policy of supplying service parts in the consecutive phases of the life cycle. In the initial phase it is important to have a close cooperation between the different departments within one company (e.g. Research & Development, Marketing & Sales). Good agreements with internal and external suppliers is essential to guarantee fast and reliable deliveries to customers. The inventory control of service parts should be centrally coordinated and competitors should be kept out of the market by means of e.g. patents and brand names. In the normal phase of the life cycle the main focus is on getting insight into the demand behavior. Using aggregate demand data and having good contacts with service organizations is essential. In the final phase the main focus is on modular design and compatible product families, such that a smooth transition to a new product is guaranteed.

Pricing

A trend can be observed in which the price of service parts itself becomes less relevant. A 'total solution concept' is offered to the customer in which the price of service parts is taken into account. The type of part (slow moving versus fast moving), type of customer (contract customer versus an once-only customer), and type of market (professional versus consumer) are factors that play an important role in the pricing policy. The customer who needs the service part tries to push back the risk as much as possible (e.g. by negotiating lease contracts or consignment inventory). The supplier who produces the service part tries to push forward the risk as much as possible (e.g. by negotiating large batch sizes and long term contracts). The parts business is often right in the middle and has to deal with both phenomena.

Activity Base Costing

The use of Activity Based Costing (ABC) in the parts business was explained by one of the PBF members. Before the introduction of ABC, costs were transferred to the various European organizations based on standard tariffs. ABC prescribes the use of cost drivers to transfer costs. In this case orderlines were used as cost driver, such that the cost of operational activities are correctly

transferred to individual products and hence to organizational units. It also proved to be a very useful instrument in the contract negotiations that were held to outsource the physical distribution activities. A general scheme for introducing ABC prescribes the following steps: identify the primary process and main (and secondary) activities within this process, determine cost drivers that are used to transfer operational costs, and finally, determine the cost targets to which the costs are transferred. The question remains how to deal with obsolescence cost when applying ABC. Since this is an important cost factor in the parts business, the applicability of ABC in the parts business is questionable.

A customer's perspective

Among the members of the PBF are also big companies that buy service parts for own usage (e.g. oil industry, power stations, railroad companies). One of these companies explained how they categorize service parts and how they treat suppliers of parts in these categories. Two discerning factors are used: value and risk. The factor 'value' contains aspects such as number of suppliers, business volume and share in supplier's turnover. The factor 'risk' contains aspects such as financial risk, HSE-risks (health, safety, and environment) and locations of suppliers. For parts with low value and low risk (the routine parts) the focus is on simplifying the ordering process. This can be achieved by applying concepts such as Electronic Data Interchange and Vendor Managed Inventory. For parts with low value and high risk (the bottleneck parts) the focus is on ensuring reliable deliveries. A good knowledge of the market situation is therefore crucial. For parts with high value and low risk (the leverage parts) the focus is on cost savings. Flexibility, coordinated purchasing, and standardization are important concepts for these parts. Finally, for parts with high value and high risk (the strategic parts) the focus is on establishing strategic alliances with suppliers. Developing what-if scenarios and negotiations with suppliers play a key role in this situation.

Distribution networks and customer service levels

The strategic goal in the parts business is to maximize customer service at minimum cost. Maximizing customer service implies that the parts inventories should be allocated as close as possible to the customer. Minimizing cost implies that the parts inventories should be centralized such that, for example, risk of obsolescence is minimized. The trade-off between these two conflicting statements is situation dependent. One of the PBF members illustrated a trend of decentralization of inventories in the past decades. Service levels are increasing, lead times are decreasing, transportation cost (especially overnight distribution) are decreasing, and the assortment is increasing. Ensuring high customer service at minimum cost will be a crucial competitive weapon in the future.

Appendix

Opportunities and threats

An increasing threat in the parts business is the rise of imitation (or second line) parts that are sold by brand independent organizations. These organizations do not offer the complete assortment of parts, but focus on the fast moving parts with high net margins. In order to compete, the original equipment maker has to adjust its prices and loses turnover. Furthermore, these imitation parts damage the brand name, complicate warranty regulations, and cause product liability claims. To deal with this threat, it is important to convince wholesalers, dealers, and importers to do business with you. Offering reasonable prices and differentiating these prices with respect to the buyer are important techniques. Under the motto 'if you cannot beat them, join them', it is also possible to offer 'white parts' to the market. Producing and selling brandless service parts can also increase revenues.

Profit center or cost center?

A strategic issue with respect to the parts business is the positioning of this business as a cost center or a profit center. The situation at one of the PBF members is that some activities are considered as cost centers (i.e., initial service tasks and warranty agreements), some as no-profit/no-loss centers (i.e., technical assistance with respect to product failures), and some as profit centers (i.e., selling service parts to the open market). In many cases companies do not sell products any more, but sell functionalities. Sales and After Sales activities have to be considered jointly. There is a focus on life time revenue: the initial product may be offered at or below cost price, as long as the service activities (including the selling of parts) are profitable. When considering the 'total solution concept' approach, failures of products should be minimized. This implies that the parts business should be minimized as well, in order to reduce the cost of providing service to the customer.

In this appendix we introduced the Parts Business Forum and its objectives. Examples of issues of present interest in the parts business were presented. In the future the Parts Business Forum will continue its periodical meetings in which these issues are discussed. The value of these meetings lies in the fact that the addressed issues emerge from the participants themselves. It therefore constitutes an important breeding ground for defining academic research in this field.

Appendix B

In this appendix we derive an exact expression for the probability $v(t)$ that the waiting time in the ERM exceeds a given time limit t (see Chapter 4, Section 4.3). Define the following stochastic variables:

- \underline{W} : Waiting time for a backordered part in the ERM
 \underline{W}_{ij} : Waiting time for a backordered part in the ERM when the system is in state (i,j) just before the arrival of a failed part

Then we have:

$$\begin{aligned} v(t) &= Pr\{ \underline{W} > t \} \\ &= \sum_{i+j \geq S} Pr\{ \underline{W}_{ij} > t \} p_{ij} \end{aligned}$$

The probability that the waiting time \underline{W}_{ij} exceeds a time period of length t is equal to the probability that at most $i+j-S$ parts are repaired in this time period. These can be parts that were in normal repair just before the arrival of a failed part, parts that were in emergency repair just before the arrival of a failed part, or parts that arrived later but were repaired during the time period of length t . Therefore:

$$\begin{aligned} Pr\{ \underline{W}_{ij} > t \} &= \sum_{x \leq i+j-S} Pr\{ x \text{ parts are repaired in time } t \} \\ &= \sum_{\substack{k+l+m \leq i+j-S \\ k \leq i, l \leq j+1}} P(i,k) Q(j+1,l) R(m) \end{aligned}$$

With:

- $P(i,k) := Pr\{ k \text{ of the } i \text{ parts are repaired in the normal mode in time period } (0,t) \}$
 $Q(j+1,l) := Pr\{ l \text{ of the } j+1 \text{ parts are repaired in the emergency mode in time period } (0,t) \}$
 $R(m) := Pr\{ m \text{ new arrivals are repaired in the emergency mode in time period } (0,t) \}$

The probability that a part that was already in repair at time 0 is repaired before time t is exponentially distributed with parameter μ (in case of normal repair) or τ (in case of emergency repair). This leads to the following expressions for $P(i,k)$ and $Q(j,l)$:

$$P(i,k) = \binom{i}{k} (1 - e^{-\tau t})^k (e^{-\tau t})^{i-k}$$

$$Q(j+1,l) = \binom{j+1}{l} (1 - e^{-\tau t})^l (e^{-\tau t})^{j+1-l}$$

The expression for $R(m)$ is more complex. The probability that m parts arrive after time 0 and are repaired before time t is equal to the probability that at least m parts arrive after time 0 and that exactly m of these parts are repaired before time t . Let q denote the probability that a part that arrives after time 0 is repaired before time t . Then we have:

$$R(m) = \sum_{j=m}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!} \binom{j}{m} (1-q)^m q^{j-m}$$

Finally, we have to find an expression for q , i.e. the probability that a part arriving after time 0 is repaired before time t . The Poisson arrivals of failed parts in the time interval $(0,t)$ are uniformly distributed (see e.g. Tijms, 1988). The probability that such a part is repaired before time t is then equal to:

$$\begin{aligned} q &= \Pr\{\text{part arriving at time } x \ (0 < x < t) \text{ is repaired before time } t\} \\ &= \Pr\{\text{repair time is smaller than } t-x \mid \text{part arrives at time } x\} \\ &= \int_0^t \frac{1 - e^{-\tau(t-x)}}{t} dx \\ &= 1 - \frac{1 - e^{-\tau t}}{\tau t} \end{aligned}$$

Appendix C

In this appendix we derive the expression given in Section 4.4.2 of Chapter 4 for the Laplace Stieltjes transform $\varphi_n(s)$ of the time T_n to bring down the number of customers from n to $n-1$ in an $M/M/1$ queue with arrival rate λ and service rate τ . Define:

- A : inter-arrival time of customers (exponentially distributed with parameter λ)
- C : cycle time between two consecutive arrivals of customers that arrive in an empty system

The cycle time C can then be written as follows:

$$C = T_1 + A \tag{C.1}$$

From Takács (1962) we can find the following expression for the Laplace-Stieltjes transform $\gamma(s)$ of the cycle time C in an $M(\lambda)|M(\tau)|\infty$ queue (page 211, expression 8):

$$\begin{aligned} \gamma(s) &= E[e^{-sC}] \\ &= 1 - \frac{s}{\lambda + s} \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=1}^k \frac{-\lambda}{s + i\tau}} \end{aligned} \tag{C.2}$$

Because T_1 and A are independent, we can write:

$$\begin{aligned} \gamma(s) &= E[e^{-sC}] \\ &= E[e^{-s(T_1 + A)}] \\ &= E[e^{-sT_1}] E[e^{-sA}] \\ &= \varphi_1(s) \frac{\lambda}{\lambda + s} \end{aligned} \tag{C.3}$$

From expression (C.2) and (C.3) we can derive the following expression for the Laplace-Stieltjes transform $\varphi_1(s)$ of T_1 :

$$\varphi_1(s) = \frac{1}{\lambda} \left\{ \lambda + s - \frac{s}{1 + \sum_{k=1}^{\infty} \prod_{i=1}^k \frac{-\lambda}{s + i\tau}} \right\} \tag{C.4}$$

Appendix

In order to calculate the Laplace-Stieltjes transform T_n for $n > 1$, note that when there are n customers in the system two things can happen. With probability $n\tau/(\lambda+n\tau)$ a customer leaves the system and with probability $\lambda/(\lambda+n\tau)$ a new customer arrives. The time that elapses till one of these two events happens is exponentially distributed with parameter $\lambda+n\tau$. Denote this time by X_n . Hence, the following relation holds:

$$\begin{aligned} T_n &= X_n && \text{with probability } \frac{n\tau}{\lambda+n\tau} \\ &= X_n + \hat{T}_{n+1} + \hat{T}_n && \text{with probability } \frac{\lambda}{\lambda+n\tau} \end{aligned}$$

where X_n , \hat{T}_{n+1} and \hat{T}_n are independent and \hat{T}_n and \hat{T}_{n+1} are stochastically identical to T_n and T_{n+1} respectively. Now we can write:

$$\begin{aligned} \varphi_n(s) &= E[e^{-sT_n}] \\ &= E[e^{-sX_n}] \left(\frac{\lambda}{\lambda+n\tau} E[e^{-s(T_{n+1}+T_n)}] + \frac{n\tau}{\lambda+n\tau} \cdot 1 \right) \\ &= E[e^{-sX_n}] \left(\frac{\lambda}{\lambda+n\tau} E[e^{-sT_{n+1}}] E[e^{-sT_n}] + \frac{n\tau}{\lambda+n\tau} \right) \\ &= \frac{\lambda+n\tau}{\lambda+n\tau+s} \left(\frac{\lambda}{\lambda+n\tau} \varphi_{n+1}(s) \varphi_n(s) + \frac{n\tau}{\lambda+n\tau} \right). \end{aligned}$$

Finally, we find ($n \geq 0$):

$$\varphi_{n+1}(s) = \frac{(\lambda+n\tau+s) \varphi_n(s) - n\tau}{\lambda \varphi_n(s)} .$$

Appendix D

In this appendix we show how to calculate the transition probabilities $\pi_{i,i-k}$ for the Emergency Repair Model when the normal repair channel has only one server. In Section 4.4.5 we derived the following expression for this situation:

$$\pi_{i,i-k} = \frac{(-\mu)^k}{k!} \frac{d^k}{ds^k} \{ \varphi_{S-x+1}(s) \}_{s=\mu}$$

In Appendix B we derived the following general recursive relation for $\varphi_{n+1}(s)$ for $n \geq 1$:

$$\varphi_{n+1}(s) = \frac{(\lambda + n\tau + s)\varphi_n(s) - n\tau}{\lambda \varphi_n(s)}$$

This expression can be rewritten as follows:

$$\varphi_n(s)(\lambda + n\tau + s - \lambda \varphi_{n+1}(s)) = n\tau \tag{D.1}$$

For ease of notation we define the following functions f and g :

$$\begin{aligned} f &:= \varphi_n(s) \\ g &:= \lambda + n\tau + s - \lambda \varphi_{n+1}(s) \end{aligned}$$

Let $f^{(k)}$, respectively $g^{(k)}$, denote the k -th derivative of f , respectively g , with respect to s . So:

$$\begin{aligned} f^{(k)} &= \varphi_n^{(k)}(s) & (k \geq 1) \\ g^{(1)} &= 1 - \lambda \varphi_{n+1}^{(1)}(s) \\ g^{(k)} &= -\lambda \varphi_{n+1}^{(k)}(s) & (k \geq 2) \end{aligned}$$

From expression (D.1) we can now derive the following expression for the k -th derivative of the product of f and g :

$$(f \cdot g)^{(k)} = \sum_{j=0}^k \binom{k}{j} f^{(j)} g^{(k-j)} = 0 \quad (k \geq 1)$$

It then follows:

$$g^{(k)} = \frac{\sum_{j=1}^k \binom{k}{j} f^{(j)} g^{(k-j)}}{-f} \tag{D.2}$$

Appendix

Expression (D.2) states that the k -th derivative of $\varphi_{n+1}(s)$ can be expressed as a function of the j -th derivative ($j < k$) of $\varphi_{n+1}(s)$ and the j -th derivative ($j \leq k$) of $\varphi_n(s)$. If we are able to derive an expression for the k -th derivative of $\varphi_1(s)$, then we can calculate the k -th derivative of $\varphi_n(s)$ for any $n > 1$, and hence we can calculate the transition probabilities $\pi_{i,k}$.

In Appendix C we derived the following expression for $\varphi_1(s)$ (see expression C.3):

$$\gamma(s) = \varphi_1(s) \frac{\lambda}{\lambda + s} \tag{D.3}$$

where $\gamma(s)$ represents the Laplace-Stieltjes transform of the cycle time in an $M(\lambda) | M(\tau) | \infty$ queue. Takács (1962, page 210, expression 2) gives the following expression for $\gamma(s)$:

$$\begin{aligned} \gamma(s) &= 1 - \frac{1}{\lambda + s} \left\{ \int_0^\infty e^{-st - \frac{\lambda}{\tau} \int_0^t \tau e^{-\tau x} dx} dt \right\}^{-1} \\ &= 1 - \frac{1}{\lambda + s} \left\{ \int_0^\infty e^{-st - \frac{\lambda}{\tau}(1 - e^{-t})} dt \right\}^{-1} \\ &= 1 - \frac{1}{\lambda + s} \left\{ e^{-\frac{\lambda}{\tau}} \int_0^\infty e^{-st} e^{\frac{\lambda}{\tau} e^{-t}} dt \right\}^{-1} \\ &= 1 - \frac{1}{\lambda + s} \left\{ e^{-\frac{\lambda}{\tau}} \int_0^\infty e^{-st} \sum_{j=0}^\infty \frac{1}{j!} \left(\frac{\lambda}{\tau} \right)^j e^{-jt} dt \right\}^{-1} \\ &= 1 - \frac{1}{\lambda + s} \left\{ e^{-\frac{\lambda}{\tau}} \sum_{j=0}^\infty \frac{1}{j!} \left(\frac{\lambda}{\tau} \right)^j \int_0^\infty e^{-(s+j\tau)t} dt \right\}^{-1} \\ &= 1 - \frac{1}{\lambda + s} \left\{ e^{-\frac{\lambda}{\tau}} \sum_{j=0}^\infty \frac{1}{j!} \left(\frac{\lambda}{\tau} \right)^j \frac{1}{s + j\tau} \right\}^{-1} \end{aligned}$$

This expression can be rewritten as follows:

$$(1 - \gamma(s)) \left(e^{-\lambda/\tau} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\lambda}{\tau} \right)^j \frac{\lambda + s}{s + j\tau} \right) = 1$$

The k -th derivative of $\gamma(s)$ can now be obtained in a similar way as before by calculating the k -th derivative of the product of f and g , where:

$$\begin{aligned} f &:= 1 - \gamma(s) \\ g &:= e^{-\lambda/\tau} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\lambda}{\tau} \right)^j \frac{\lambda + s}{s + j\tau} \end{aligned}$$

and

$$\begin{aligned} f^{(k)} &= -\gamma^{(k)}(s) \\ g^{(k)} &= e^{-\lambda/\tau} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\lambda}{\tau} \right)^j \frac{(\lambda - j\tau)(-1)^k k!}{(s + j\tau)^k} \end{aligned}$$

Once we have obtained the expression for $\gamma^{(k)}(s)$, we can calculate the k -th derivative of $\phi_1(s)$ as follows (see expression D.3):

$$\phi_1^{(k)}(s) = \gamma^{(k)}(s) \left(1 + \frac{s}{\lambda} \right) + \frac{k\gamma^{(k-1)}(s)}{\lambda}$$

Appendix E

In this appendix we present the cost and lead time characteristics of the emergency supply structure for the case study in Chapter 5, Section 5.8. The cost of a lateral transshipment and the cost of a direct delivery consist of a fixed cost plus a variable cost per kilogram of shipped weight. Table E.1 shows the cost of direct deliveries from the central warehouse to the various national warehouses for two situations: delivery time of 16 hours and delivery time of 68 hours. The cost parameters are normalized.

| Cost | Fi | Fr | Ge | No | Sp | Sw | UK |
|---------------------|-----|-----|-----|-----|-----|-----|-----|
| Fixed | 10 | 6 | 3.8 | 6.8 | 6 | 6 | 15 |
| Variable (16 hours) | 1.7 | 1.0 | 0.5 | 1.5 | 1.2 | 1.2 | 0.5 |
| Variable (68 hours) | 0.6 | 0.3 | 0.2 | 0.5 | 0.2 | 0.2 | 0.1 |

Table E.1: Fixed and variable cost (per kg.) for direct shipments from the central warehouse.

Table E.2 shows the cost of using lateral transshipments between national warehouses. Again, the cost parameters are normalized and two scenarios are presented: delivery time of 16 hours and delivery time of 68 hours.

| | | Fr | Ge | No | Sp | Sw | UK |
|----|------|-----|-----|-----|-----|-----|-----|
| Fi | F | 17 | 14 | 18 | 18 | 18 | 32 |
| | V-16 | 3.5 | 2.4 | 1.8 | 3.9 | 1.0 | 2.8 |
| | V-68 | 0.8 | 0.3 | 0.8 | 2.2 | 0.4 | 0.5 |
| Fr | F | | 3.7 | 7.0 | 6.1 | 8.2 | 15 |
| | V-16 | | 0.7 | 1.9 | 1.7 | 1.2 | 1.2 |
| | V-68 | | 0.1 | 0.7 | 0.2 | 0.2 | 0.2 |
| Ge | F | | | 4.8 | 3.9 | 6.2 | 17 |
| | V-16 | | | 1.4 | 1.6 | 0.8 | 0.6 |
| | V-68 | | | 0.5 | 0.2 | 0.2 | 0.3 |
| No | F | | | | 6.9 | 8.1 | 19 |
| | V-16 | | | | 2.7 | 0.5 | 1.9 |
| | V-68 | | | | 2.0 | 0.3 | 0.3 |
| Sp | F | | | | | 6.1 | 16 |
| | V-16 | | | | | 1.3 | 1.2 |
| | V-68 | | | | | 0.9 | 0.3 |
| Sw | F | | | | | | 25 |
| | V-16 | | | | | | 1.4 |
| | V-68 | | | | | | 0.3 |

Table E.2: Fixed and variable cost (per kg.) for lateral transshipments between national warehouses.

F : Fixed cost for a lateral transshipment.

V-16 : Variable cost per kilogram for a lateral transshipment with a lead time of 16 hours.

V-68 : Variable cost per kilogram for a lateral transshipment with a lead time of 68 hours.

Appendix F

In this appendix we explain the techniques that were used to determine the length of the transient phase and the length of the subruns in the simulation experiment presented in Chapter 6.

Transient phase

We are interested in the steady state performance of the supply system for various flexibility policies. The time that expires until the system operates in steady state under simulation is called the transient phase. The length of this 'warm-up' period in the simulation is determined graphically, according to Welch (see Law and Kelton, 1991). We express the length of the transient phase in average number of customer arrivals at each local warehouse. An example is presented in figure F.1, where the total cost performance is presented as a function of the length of the transient phase (case 5, policy 12). From this figure it is clear that the average cost performance does not change dramatically anymore after a certain time period.

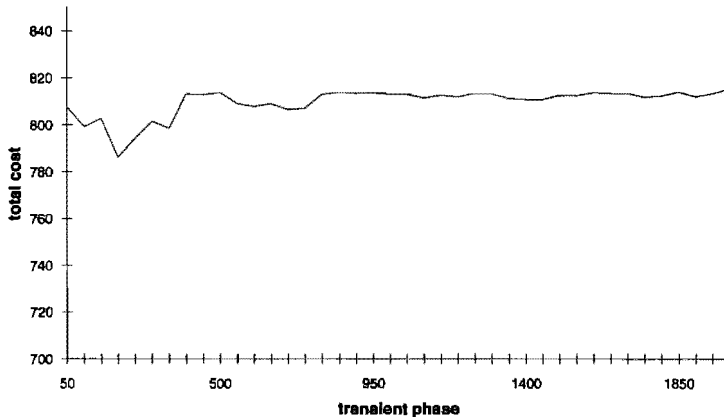


Figure F.1: Cost behavior for different lengths of the transient phase

Applying this graphical technique to several scenarios (i.e. combinations of flexibility policy and parameter setting), resulted in a length of the transient phase that equals the arrival of approximately 2000 customers at each local warehouse.

Subrun length

In order to calculate confidence intervals for the performance measures under consideration, we determined ten replications for each performance measure in each scenario. For this reason, we divided the simulation period in ten equal periods, so-called subruns. The length of each subrun must be long enough to guarantee independence between subsequent realizations of a performance indicator. To determine the length of a subrun period, we applied the Von Neumann statistic q :

$$q = \frac{\sum_{i=2}^N (x_i - x_{i-1})^2}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

with:

- N = number of subruns
- x_i = realization of performance indicator in subrun i ($i=1..N$)
- \bar{x} = average value of performance indicator over all subruns

The Von Neumann statistic q is normally distributed with mean $\mu=2$ and variance $\sigma^2=4(N-2)/(N^2-1)$. The subrun length is long enough to guarantee independence if $|q - 2| < z_{\alpha/2} \sigma$. The parameter $z_{\alpha/2}$ follows from the expression:

$$Pr \{ |X| < z_{\alpha/2} \} = \alpha$$

where X is a stochastic variable with a standard normal distribution. Application of this technique in our simulation experiment with a significance level of 95% (i.e. $\alpha = 0.95$), resulted in a subrun length that equals the arrival of approximately 1000 customers at every local warehouse.

Appendix G

In Section 6.6.1 we investigate the effect of using pipeline flexibility. We compare the total cost, average response time, and fill rate performance of policies 1 to 6 (i.e., policies without pipeline flexibility) with policies 7 to 12 (i.e., policies with pipeline flexibility). Analysis Of Variance (ANOVA) is applied to test the hypothesis that the average performance (in terms of total cost, response time, and fill rate) is significantly different as a result of using pipeline flexibility. The F-values, representing the ratio of the *between-group variance* (i.e., variance as a result of adding pipeline flexibility to policies 1 to 6) and the *within-group variance* (i.e., variance as a result of randomness in simulation for a given policy), are presented in tables G.1 (total cost), G.2 (response time), and G.3 (fill rate). If the F-value exceeds the critical value 4.41, the performance of the policy with pipeline flexibility (policies 1 to 6) is significantly different from the performance of the policy without pipeline flexibility (policies 7 to 12).

| Case | Policy comparison | | | | | |
|------|-------------------|-------|-------|-------|-------|-------|
| | 1-7 | 2-8 | 3-9 | 4-10 | 5-11 | 6-12 |
| 1 | 6.44 | 0.07 | 6.66 | 5.89 | 0.07 | 0.07 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.47 | 0.05 | 0.47 | 0.47 | 0.05 | 0.05 |
| 4 | 37.00 | 0.15 | 7.97 | 7.47 | 0.15 | 0.15 |
| 5 | 684.2 | 15.17 | 51.24 | 42.84 | 14.24 | 11.36 |
| 6 | 18.07 | 0.07 | 17.62 | 15.94 | 0.07 | 0.07 |
| 7 | 33.40 | 0.12 | 26.81 | 23.87 | 0.12 | 0.12 |
| 8 | 1.06 | 0.07 | 1.06 | 1.06 | 0.07 | 0.07 |

Table G.1: F-values resulting from ANOVA for total cost performance.

| Case | Policy comparison | | | | | |
|------|-------------------|-------|-------|-------|-------|-------|
| | 1-7 | 2-8 | 3-9 | 4-10 | 5-11 | 6-12 |
| 1 | 550.9 | 0.25 | 485.0 | 342.3 | 0.25 | 0.25 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 8.69 | 1.37 | 8.69 | 8.69 | 1.37 | 1.37 |
| 4 | 1146 | 1.15 | 43.73 | 42.41 | 1.15 | 1.15 |
| 5 | 8457 | 13.23 | 279.1 | 289.1 | 16.67 | 10.24 |
| 6 | 321.2 | 0.38 | 317.4 | 211.4 | 0.38 | 0.38 |
| 7 | 1052 | 0.78 | 294.9 | 281.5 | 0.78 | 0.78 |
| 8 | 13.62 | 1.55 | 13.62 | 13.62 | 1.55 | 1.55 |

Table G.2: F-values resulting from ANOVA for response time performance.

| Case | Policy comparison | | | | | |
|------|-------------------|-------|-------|-------|-------|-------|
| | 1-7 | 2-8 | 3-9 | 4-10 | 5-11 | 6-12 |
| 1 | 6.44 | 0.02 | 6.44 | 5.41 | 0.02 | 0.02 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.90 | 0.04 | 0.19 | 0.25 | 0.04 | 0.04 |
| 5 | 624.1 | 35.37 | 240.9 | 217.2 | 27.68 | 23.48 |
| 6 | 0.10 | 0 | 0.10 | 0.10 | 0 | 0 |
| 7 | 3.47 | 1.03 | 3.47 | 3.47 | 1.03 | 1.03 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 |

Table G.3: F-values resulting from ANOVA for fill rate performance.

In Section 6.6.2 we identified for each case the cost-minimal policy. We apply ANOVA for two cases to test if the differences in performance is statistically significant. In case 2 we compare the performance of policy 3 (or, equivalently, policy 5) with the performance of policy 4 (or, equivalently, policy 6). In case 5 we compare the performance of policies 2, 5, and 6. The resulting F-values are presented in table G.4. The critical F-value is 4.41 in all situations.

| Case | Comparison | TC | RT | FR |
|------|-----------------------|------|-------|------|
| 2 | Policy 3 - Policy 4 : | 3.15 | 25.92 | - |
| 5 | Policy 2 - Policy 5 : | 0.18 | 0.14 | 0.01 |
| | Policy 2 - Policy 6 : | 1.13 | 1.91 | 0.12 |
| | Policy 5 - Policy 6 : | 0.45 | 1.06 | 0.06 |

Table G.4: F-values resulting from ANOVA with respect to Total Cost (TC), Response Time (RT) and Fill Rate (FR) performance.

SUMMARY

The central theme of this book is the use of flexibility in complex repair- and distribution networks for service parts. These parts are needed to restore failures of technical systems or equipment at customer sites. Since customers are often scattered over a large geographical area, multi-echelon distribution networks are needed to ensure a fast response in case of failures. Such a distribution system generally consists of a central warehouse that supplies national warehouses in different countries. These national warehouses in their turn supply regional or local warehouses within a country. Service engineers collect service parts at these warehouses at the lowest echelon in the system to perform repair jobs at customer sites. Sometimes the engineers have an assortment of service parts in their car, the so-called car stocks, that can also be considered an inventory level in the distribution system. Many service parts are characterized by a high value and a low demand frequency. Therefore, it is economically interesting to repair these parts after failure. The return flow of failed parts to local or central repair shops, and the repair activities that are performed in these shops, constitute a complex network in itself. The design and control of the distribution and repair processes of service parts are topic of research in this book.

Two important trends can be identified in the service part business in the last decade. First, reducing costs is an important issue. The use of multi-layered distribution networks for supplying service parts to the customer, has led to high inventories of service parts throughout the network. The associated holding cost and invested capital is an important cost factor. Therefore, many companies focus on reducing inventories in the supply system. Second, increasing the service performance towards the customer has also become a strategic issue. Higher service performance, in terms of e.g. lower customer response times and higher fill rates, is nowadays a necessary condition to survive.

A way to decrease cost and increase service is the use of flexibility. In this thesis we present and analyze several flexibility policies that can be used to meet this goal. We present a framework for control, the Service Part Supply System, in which we identify six flexibility opportunities. A distinction is made between repair flexibility and supply flexibility. Repair flexibility can be applied to increase the performance (i.e. lower cost and higher service) of the repair processes of failed parts. The following three options are proposed and discussed:

- 1) Return Flow Flexibility
- 2) Work Order Release Flexibility
- 3) Repair Shop Flexibility

Return Flow Flexibility concerns the organization and control of return flows of failed service parts from the field. Work Order Release Flexibility concerns the release of failed parts to the repair

shops for repair (what type of parts, how many of each type, and where to repair these parts). Repair Shop Flexibility concerns the use of various types of flexibility within the repair shop (e.g. scheduling techniques and flexible manpower).

Supply flexibility can be applied to increase the performance of the distribution processes of service parts throughout the network. The following three options are proposed and discussed:

- 4) Allocation Flexibility
- 5) Pooling Flexibility
- 6) Direct Shipment Flexibility

Allocation Flexibility concerns the assortment decision (what type of parts to stock at an inventory location in the network), the stock level decision (how many of each part to stock), and the shortage allocation decision (how to divide inventory in case of shortages). Pooling Flexibility concerns the sharing of inventory between inventory locations at the same level in the supply system (e.g. national, regional, or local) in case of stock-out situations. Direct Shipment Flexibility concerns the use of direct deliveries of parts from a higher inventory level in the system in case of stock-out situations.

Implementation of any kind of flexibility involves an economic trade-off: what are the costs (e.g. investment in information technology, use of special delivery couriers) and what are the benefits (e.g. lower response times, inventory reduction) of implementing flexibility. In this book we present two analytical models that support the economic trade-off for implementing flexibility.

The first model, the Emergency Repair Model, considers an inventory location at which failed parts arrive. These parts are exchanged for serviceable parts from stock on hand and the failed part is sent into repair. In case of a stock-out situation a backorder is created and penalty costs are incurred. Two alternative repair modes are available: normal repair (cheap but with a long repair lead time) and emergency repair (expensive but with a short repair lead time). The emergency repair mode is applied when, upon arrival of a failed part at the inventory location, the net inventory of serviceable parts is lower than a so-called emergency trigger level. Otherwise, the normal repair mode is applied. We present an exact analysis of the operating characteristics of this system. Given the process parameters (demand rate, normal repair rate, and emergency repair rate) and the cost parameters (cost of holding inventory, penalty cost for backorders, normal repair cost, and emergency repair cost), we can calculate the initial stock level and the emergency trigger level that minimize the total cost function. The numerical results indicate that the emergency trigger level is an important decision variable. Although in practice one often resorts to emergency procedures when the inventory on hand drops to zero (i.e. emergency trigger level is equal to zero), we show that

Summary

positive or negative trigger levels can be cost-optimal as well. Another important observation is that different types of service constraints (e.g. minimum fill rate or maximum response time) can result in different values for the initial stock level and the emergency trigger level that are cost-optimal. An important insensitivity result that we obtained by using simulation is the fact that the service performance (especially the fill rate performance) is to a large extent independent of the choice of repair lead time distribution, but only depends on the average repair lead time. However, the assumption of Poisson demand turned out to be very essential. This means that if demand is less erratic (e.g. preventive maintenance is applied) the model should be interpreted with care. Finally, we compared the performance of the policy prescribed by the Emergency Repair Model with the following policies: 1) emergency repair is not possible; 2) trigger level is equal to zero; and 3) service parts that are repaired in the normal mode are not used to satisfy backorders. The results show that the policy prescribed by the Emergency Repair Model performs significantly better in many cases, depending on the parameter setting of the situation at hand.

The second model, the Emergency Supply Model, considers a two-echelon inventory system for service parts. Several local warehouses are supplied from one central warehouse, which in its turn is supplied from a production plant with infinite capacity. In case of a demand for a service part at a local warehouse, the following flexibility policy is applied: 1) satisfy the demand from stock on hand and issue a replenishment order to the central warehouse; 2) when the local warehouse is out of stock, a lateral transshipment from any other local warehouse with stock on hand is issued (i.e. pooling flexibility is applied); 3) when all local warehouses are out of stock, a direct delivery from the central warehouse is applied (i.e. direct shipment flexibility is applied); and 4) when also the central warehouse is out of stock, a direct shipment from the production plant is applied (which is always possible). The Emergency Supply Model is an approximate model but simulation is used to demonstrate the accuracy of the model performance. We compare the cost performance of the Emergency Supply Model with the cost performance of a model in which no flexibility is applied at all (i.e. customers that arrive in a stock-out situation have to wait for a normal replenishment order to arrive). The results show that cost savings up to 44% can be obtained when applying the flexibility policy as prescribed by the Emergency Supply Model. Again we also find an important insensitivity result: the model performance is to a large extent independent of the choice of the distribution of the shipment times. With respect to the use of pooling flexibility, we compared the performance of different pooling concepts. For reason of analytical tractability, we have to assume complete pooling, i.e., all local warehouses act as potential source for lateral shipments in case of a stock-out situation. We simulated two alternative pooling concepts: fixed pooling (local warehouses are divided into pooling groups that share inventory) and variable pooling (a local warehouse can only make a request for a lateral transshipment at its' neighbouring warehouses). The results show that the difference between fixed and variable pooling is negligible. The differences with the complete pooling concept are only significant for low fill rates. In case of high fill rates

(> 70%), the differences between the three concepts are negligible.

In this book we also present a simulation study in which different flexibility policies in a three-echelon inventory system are evaluated. A specific flexibility policy consists of a selection of the following flexibility modules: 1) pipeline flexibility (the arrival time of outstanding orders is checked before applying other types of flexibility); 2) pooling flexibility (sharing of inventory between local warehouses in the same country); 3) direct shipment flexibility from the national warehouse, the central warehouse, or the production plant. The simulation results show that the use of pooling flexibility is optimal. If pooling flexibility is allowed, other flexibility options offer no additional added value in terms of cost reduction. If pooling flexibility is not allowed, direct shipment flexibility from the national warehouse is optimal. The use of pipeline flexibility only proves to be interesting when pooling flexibility is not allowed.

This book shows that implementation of flexibility in Service Part Supply Systems can be very rewarding. The economic trade-off that is associated with the implementation of any kind of flexibility should take into consideration the service requirements and the relevant cost factors that are influenced by this decision. This trade-off is dependent on situational aspects, such as type of product, type of customer, and consequences of stock-out situations. Flexibility is a necessary condition to meet the competitive challenges of decreasing costs, and at the same time, increasing service performance.

SAMENVATTING

Het centrale thema van dit boek is het gebruik van flexibiliteit in complexe reparatie- en distributienetwerken voor reserve onderdelen. Deze onderdelen zijn nodig om storingen op te lossen aan technische systemen die bij de klant staan opgesteld. Omdat klanten vaak geografisch enorm verspreid zijn, is er een behoefte aan multi-echelon distributienetwerken zodat een snelle respons naar alle klanten gegarandeerd kan worden. Zo'n distributienetwerk bestaat meestal uit een centraal magazijn dat nationale magazijnen in verschillende landen bevoorraadt. Van daar uit worden regionale of lokale magazijnen binnen de verschillende landen voorzien van onderdelen. Service monteurs halen hier onderdelen op die ze nodig hebben om reparaties bij de klant uit te voeren. Soms hebben deze monteurs zelf ook een eigen autovoorraad die ook als een voorraadniveau in het netwerk gezien kan worden. Een generiek kenmerk van veel onderdelen is een hoge prijs en een lage vraag. Dit betekent dat het economisch interessant is om deze onderdelen, als ze defect zijn, te repareren. De retourstroom van defecte onderdelen naar de reparatiecentra en de reparatieactiviteiten die in deze centra worden uitgevoerd, vormen op zich weer een complex netwerk. Het ontwerp en de beheersing van distributie- en reparatieprocessen van service onderdelen zijn onderwerp van onderzoek in dit boek.

In het laatste decennium kunnen een tweetal belangrijke trends worden waargenomen. Enerzijds is kostenbeheersing een belangrijk thema geworden. Het gebruik van multi-echelon distributienetwerken voor service onderdelen heeft geleid tot hoge voorraden. De bijbehorende voorraadkosten en het hoge geïnvesteerd vermogen vormen een belangrijke kostenfactor. Veel bedrijven zijn daarom bezig met het reduceren van deze voorraden in het distributienetwerk. Anderzijds wordt het strategische belang van een toenemende service prestatie naar de klant onderkend. Hogere service (b.v. lagere responstijden en een toenemende fractie van de vraag die direct uit voorraad geleverd kan worden) is dan ook een noodzakelijke voorwaarde om te kunnen overleven.

Gebruik van flexibiliteit kan een middel zijn om kosten te verlagen en service te verhogen. In dit proefschrift presenteren en analyseren we verschillende flexibiliteitsopties die hiervoor gebruikt kunnen worden. We presenteren een raamwerk, het 'Service Part Supply System', waarin we zes verschillende flexibiliteitsopties identificeren. Er wordt daarbij onderscheid gemaakt tussen reparatieflexibiliteit en distributieflexibiliteit. Reparatieflexibiliteit heeft betrekking op de operationele beheersing van de reparatieprocessen van defecte onderdelen. We onderkennen de volgende drie opties:

- 1) Retourstroom Flexibiliteit
- 2) Werkordervrijgave Flexibiliteit

3) Reparatiecentrum Flexibiliteit

'Retourstroom Flexibiliteit' heeft betrekking op de organisatie en beheersing van de retourstromen van defecte onderdelen uit het veld. 'Werkordervrijgave Flexibiliteit' heeft betrekking op de vrijgave van defecte onderdelen voor reparatie (welke onderdelen, hoeveel onderdelen, en waar moeten ze worden gerepareerd). 'Reparatiecentrum Flexibiliteit' heeft betrekking op verschillende vormen van flexibiliteit gedurende het reparatieproces (b.v. prioriteitsregels en variabele werktijden).

Distributieflexibiliteit heeft betrekking op de operationele beheersing van de distributieprocessen van onderdelen in het netwerk. We onderkennen wederom drie opties:

- 4) Allocatie Flexibiliteit
- 5) Pooling Flexibiliteit
- 6) Direct Leveren Flexibiliteit

'Allocatie Flexibiliteit' heeft betrekking op de assortimentsbepaling (welke onderdelen moeten waar op voorraad worden gehouden?), de voorraadniveau's (hoeveel moet op voorraad worden gehouden?) en de allocatieregel bij tekorten (hoe moet de voorraad worden gealloceerd als er een tekort is?). 'Pooling Flexibiliteit' heeft betrekking op het delen van voorraad met andere magazijnen in dezelfde laag van het distributienetwerk (b.v. nationaal, regionaal, of lokaal) in geval van tekorten. 'Direct Leveren Flexibiliteit' heeft betrekking op het gebruik maken van directe zendingen van onderdelen vanaf een hoger niveau in het netwerk in geval van tekorten.

Implementatie van flexibiliteit gaat altijd gepaard met een economische afweging: wat zijn de kosten (b.v. investeringen in informatietechnologie, gebruik van koeriersdiensten) en wat zijn de opbrengsten (b.v. lagere responstijden, voorraadbeparingen) van gebruik van flexibiliteit. In dit boek presenteren we twee analytische modellen die deze economische afweging bij het gebruik van flexibiliteit ondersteunen.

Het eerste model, het 'Emergency Repair Model', modelleert een voorraadpunt waar defecte onderdelen arriveren. Een defect onderdeel wordt vervangen door een reserve onderdeel uit voorraad en wordt vervolgens gerepareerd. Als er geen reserve onderdeel beschikbaar is, wordt er een nabestelling gegenereerd en moeten er boetekosten worden betaald. Er zijn twee onafhankelijke reparatiekanalen: normale reparatie (goedkoop maar met een lange doorlooptijd) en spoedreparatie (duur maar met een korte doorlooptijd). De spoedreparatie wordt uitgevoerd als de netto voorraad, bij aankomst van een defect onderdeel, lager is dan een bepaalde drempelwaarde. Zo niet, dan wordt een normale reparatie uitgevoerd. We presenteren een exacte analyse van het model. Voor gegeven procesparameters (vraagfrequentie en reparatiesnelheden) en kostenparameters (voorraadkosten,

Samenvatting

reparatiekosten en boetekosten) berekenen we de initiële voorraadhoogte (van reserve onderdelen) en de drempelwaarde (voor spoedreparaties) waarbij de totale kosten minimaal zijn. De numerieke resultaten tonen aan dat deze drempelwaarde een belangrijke beslisvariabele is. In de praktijk maakt men vaak gebruik van spoedprocedures als de fysieke voorraad nul is (d.w.z. de drempelwaarde is gelijk aan nul). Ons model toont aan dat een positieve of een negatieve drempelwaarde vaak beter is. Verder is het zo dat deze waarden afhankelijk zijn van de service eisen die worden gesteld (b.v. minimale 'fill rate' of maximale responstijd). Een belangrijk ongevoeligheidsresultaat dat volgt uit de numerieke experimenten is het feit dat de serviceprestatie (met name de 'fill rate') vrijwel onafhankelijk is van de kansverdeling van de reparatietijden en alleen afhangt van de gemiddelde reparatietijd. De aanname van een Poisson verdeelde vraag blijkt echter wel essentieel. Dus bij een beter voorspelbare vraag (b.v. in het geval van preventief onderhoud) moet dit model zeer voorzichtig gehanteerd worden. Tenslotte hebben we de beheersingsregel van het 'Emergency Repair Model' vergeleken met enkele andere regels: 1) spoedreparatie is geen optie; 2) drempelwaarde is gelijk aan nul; 3) onderdelen die uit het normale reparatieproces komen, mogen niet worden gebruikt voor naleveringen. Uit de resultaten blijkt dat de regel van het 'Emergency Repair Model' in veel gevallen tot grote kostenbesparingen kan leiden t.o.v. de drie alternatieve regels.

Het tweede model, het 'Emergency Supply Model', modelleert een 2-echelon voorraadstelsel voor reserve onderdelen. Meerdere lokale magazijnen worden bevoorraadt door een centraal magazijn, dat op zijn beurt weer bevoorraadt wordt door een fabriek met oneindige capaciteit. Als er bij een lokaal magazijn een vraag ontstaat naar een onderdeel, dan wordt de volgende flexibiliteitsregel toegepast: 1) indien mogelijk wordt de vraag uit voorraad geleverd en wordt een aanvulorder bij het centrale magazijn geplaatst; 2) als dit niet mogelijk is, dan vindt er een laterale zending vanaf een willekeurig ander lokaal magazijn (met voorraad) plaats; 3) als alle lokale magazijnen zonder voorraad zitten, dan vindt er een directe levering vanaf het centrale magazijn plaats; en 4) als ook dit centrale magazijn geen voorraad meer heeft, dan vindt er een directe levering vanaf de fabriek plaats (wat altijd mogelijk is). Het 'Emergency Supply Model' is een approximatief model, maar simulatie toont aan dat het model zeer nauwkeurig is. We vergelijken de operationele kosten van dit model met de operationele kosten van een ander model waarin geen enkele vorm van flexibiliteit mogelijk is (d.w.z. klanten die in een buitenvoorraad situatie arriveren, moeten wachten op een naleveringsorder van het centrale magazijn). De resultaten tonen aan dat kostenbesparingen tot 44% mogelijk zijn als we de flexibiliteitsregel uit het 'Emergency Supply Model' hanteren. Ook bij dit model vinden we weer een belangrijk ongevoeligheidsresultaat: de prestatie van het model is in grote mate onafhankelijk van de kansverdeling van de transporttijden. Verder hebben we verschillende pooling-concepten vergeleken. De analyse van het model is gebaseerd op de aanname van volledige pooling. Dat wil zeggen dat ieder lokaal magazijn een potentiële bron is voor laterale zendingen naar ieder ander lokaal magazijn. Met behulp van simulatie evalueren we de prestatie van twee alternatieve pooling-concepten: vaste pooling (lokale magazijnen zijn verdeeld in vaste groepen

en pooling vindt alleen plaats tussen magazijnen in dezelfde groep) en variabele pooling (lokale magazijnen kunnen alleen een beroep doen op hun buurmagazijnen voor laterale zendingen). De resultaten tonen aan dat het verschil tussen vaste en variabele pooling verwaarloosbaar is. De verschillen in prestatie in vergelijking met volledige pooling is alleen significant bij hoge 'fill rates'. Bij lage 'fill rates' (<70%) zijn de verschillen tussen de drie concepten verwaarloosbaar.

In dit boek presenteren we ook een simulatiestudie waarin we verschillende flexibiliteitsregels in een 3-echelon voorraadstelsel evalueren. Een bepaalde flexibiliteitsregel bestaat uit een aantal van de volgende flexibiliteitsmodules: 1) pijplijnflexibiliteit (de verwachte aankomsttijd van uitstaande aanvulorders wordt vergeleken met andere flexibiliteitsopties); 2) poolingflexibiliteit (delen van voorraad met andere magazijnen in hetzelfde land); 3) direct leveren flexibiliteit vanaf de landenmagazijnen, het centrale magazijn, of de fabriek. De simulatieresultaten tonen aan dat het gebruik van poolingflexibiliteit optimaal is. Als poolingflexibiliteit een optie is, dan dragen andere flexibiliteitsmodules niet significant bij aan verdere kostenverlaging. Als poolingflexibiliteit geen optie is, dan is het optimaal om gebruik te maken van directe leveringen vanaf de landenmagazijnen. Het toepassen van pijplijnflexibiliteit is alleen interessant als poolingflexibiliteit niet toegestaan is.

Dit boek toont aan dat implementatie van flexibiliteit in 'Service Part Supply Systems' zeer interessant kan zijn. De economische afweging die gepaard gaat met de implementatie van flexibiliteit moet rekening houden met de service-eisen en de relevante kostenfactoren, die door deze beslissing beïnvloed worden. Deze afweging is afhankelijk van situationele factoren, zoals het type produkt, soort klant en gevolgen van buitenvoorraadsituaties. Flexibiliteit is een noodzakelijke voorwaarde om kosten te verlagen en tegelijkertijd service te verhogen. Het is een middel om de toenemende concurrentie het hoofd te bieden en zelf een voorsprong te nemen.

CURRICULUM VITAE

The author of this book was born on June 8, 1968 in Well. In 1986 he received his high school diploma from the "Jeruzalem College" in Venray, after which he started his study Technical Mathematics at the Eindhoven University of Technology. He received his Master's Degree in 1992 after a research project on production and inventory control in multi-echelon distribution networks. After his graduation he started a PhD research project at the Graduate School of Industrial Engineering and Management Science at the Eindhoven University of Technology. The project concerned the logistical control of complex distribution systems for service parts and was supervised by prof.dr. Ton de Kok. During his research he was involved with several projects in industry. He also spent some time at the Royal Institute of Technology in Stockholm, Sweden. This dissertation concludes his research. From January 1997 the author is employed as a researcher at Baan Company where he is involved with the development of Enterprise Resource Planning information systems.

STELLINGEN

behorende bij het proefschrift

DESIGN AND CONTROL OF SERVICE PART DISTRIBUTION SYSTEMS

van

Jos Verrijdt

I

Als gebruik gemaakt kan worden van 'pooling flexibility' in een distributienetwerk voor service parts, dan heeft implementatie van andere vormen van flexibiliteit weinig zin.

(Dit proefschrift)

II

Als alleen de 'fill rate' als prestatie-indicator wordt gebruikt in distributienetwerken voor service parts, dan zou toepassing van flexibiliteit altijd onwenselijk zijn.

(Dit proefschrift)

III

De meeste voorraadmodellen voor service parts veronderstellen een Poisson verdeelde vraag. Validatie van deze aanname is essentieel bij het toepassen van dit soort modellen in de praktijk.

(Dit proefschrift)

IV

Kwantitatieve modelanalyses van bedrijfsproblemen kent twee extreme varianten: wiskundige modelanalyse en computersimulatie. Voor een echte doorgronding van wetenschappelijke onderzoeksvragen dient eerst wiskundig analytisch onderzoek te worden verricht. Simulatie dient alleen toegepast te worden als ondersteuning van en/of aanvulling op analytisch onderzoek.

V

Samenwerking in onderzoek is een noodzakelijke voorwaarde. Elke promovendus zou daarom verplicht moeten worden om een deel van zijn of haar onderzoekstijd aan een andere (buitenlandse) onderzoeksinstelling door te brengen.

VI

De kwaliteit van het NOS-journaal vertoont een omgekeerd evenredig verband met het aantal Nederlandstalige commerciële TV-zenders.

VII

De grote partijen in de Nederlandse politiek vertegenwoordigen uiteenlopende ideologische principes. Eenmaal gezeten op het regeringspluche is hier weinig meer van te merken.

VIII

Het aanbieden van een pluspakket met meerdere TV-zenders door kabelexploitanten is een verkapte vorm van subsidiëring van onrendabele zenders.

IX

Als de politiek de agenda van het wetenschappelijk onderzoek gaat bepalen, dan verdient dit onderzoek niet langer het predikaat 'onafhankelijk'.

X

We leven in Nederland in een democratische samenleving. Het niet hebben van de doodstraf in Nederland bewijst dat het democratisch principe echter niet in alle gevallen consequent toegepast moet worden.

XI

Tweede kamerleden die voor het houden van referenda zijn, ontlopen hun eigen verantwoordelijkheid.

XII

Het vieren van carnaval verhoogt het relativiseringsvermogen.