Spare Parts Planning and Control for Maintenance Operations

This thesis is number D175 of the thesis series of the Beta Research School for Operations Management and Logistics. The Beta Research School is a joint effort of the School of Industrial Engineering and the department of Mathematics and Computer Science at Eindhoven University of Technology, and the Center for Production, Logistics and Operations Management at the University of Twente.

A catalogue record is available from the Eindhoven University of Technology Library.

ISBN: 978-90-386-3475-3

Printed by Proefschriftmaken.nl || Uitgeverij BOXPress
Cover Design by Roy Lurken - BureauNobel

This research has been funded by NedTrain.

Spare Parts Planning and Control for Maintenance Operations

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus, prof.dr.ir. C.J. van Duijn, voor een commissie aangewezen door het College voor Promoties in het openbaar te verdedigen op maandag 11 november 2013 om 16.00 uur

door

Joachim Jacob Arts

geboren te Eindhoven

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof.dr. A.G.L. Romme
1^{e} promotor:	prof.dr.ir. G.J.J.A.N. van Houtum
2^{e} promotor:	prof.dr. A.G. de Kok
leden:	prof.dr.ir. O.J. Boxma
	prof.dr.ir. L.A.M. van Dongen (Universiteit Twente)
	prof.dr. R. Levi (Massachusetts Institute of Technology)
	prof.dr. S. Minner (Technische Universität München)
	prof.dr. C. Witteveen (Technische Universiteit Delft)

Acknowledgments

Although this thesis has only one author, it is in fact the result of contributions by many people.

First I would like to thank my first *promotor* Geert-Jan van Houtum for his supervision and support. He taught me the tricks of the research trade as well as how to navigate the academic world. He made sure all the research fitted in a larger perspective while also helping out with details and technicalities.

I would like to thank NedTrain for funding my PhD research. Working in the maintenance development department of NedTrain under the supervision of Bob Huisman was a great pleasure. Bob Huisman created the perfect environment where practice and research could meet. Apart from being involved with the particulars of research, he also has a keen eye for the human aspect of the research endeavor. I also thank Michel Wilson and Jorge Parada for our joint discussions on research and applications at NedTrain. I thank Leo van Dongen for leading this collaboration between NedTrain and academia, and for chairing the steering committee of this collaboration. I also thank him for his input during meetings of the steering committee and for being on my thesis committee.

At NedTrain, I am also indebted to Joost Florie and Guido Aerts for their input on modeling issues and for acquainting me with the particulars of NedTrain's supply chain and repair shop operations.

I thank Maarten Driessen for our collaboration on Chapter 2 of this thesis. I enjoyed our many conversations on how to apply scientific knowledge in practice. I also thank Kristina Sharypova for tolerating such lengthy conversations in our office, bringing a pleasant atmosphere to the office, and reminding us of the more important things in life.

I would like to thank Simme Douwe Flapper for our collaboration on Chapter 3. Chapter 3 has also benefited from the graduation projects of Karin Vernooij, Anne Basten, and Martine Rousseau.

I thank Rob Basten for being my daily supervisor when he was in Eindhoven and his contributions to Chapter 4 of this thesis. Rob Basten, together with Frank Karsten and Willem van Jaarsveld, have provided the necessary feedback and reflection on academic life.

Chapter 5 of this thesis has benefited from the master thesis projects of Martijn van Aspert and Nadine Loeffen.

I would like to thank Retsef Levi for hosting my visit to MIT and our joint work

on Chapter 6 of this thesis. In his busy schedule, he managed to free time for our collaboration and he was always very sharp during our meetings. Working at the Operations Research Center at MIT was truly stimulating. I would also like to thank my sister Saskia and her husband John for letting me and my family stay at their home and otherwise making our stay in the greater Boston area a pleasant experience. My visit to MIT was partially funded by the Prins Bernhard Fellowship supported by the De Breed Kreiken innovation fund.

Ton de Kok became my second *promotor* somewhat late in my PhD project, but has shown an active interest in my work from the beginning. I really enjoyed our many conversations on very diverse topics.

I thank Onno Boxma, Stefan Minner, and Cees Witteveen for being on my thesis committee and their valuable feedback on my work.

Qiushi Zhu was my office mate for almost three years. We had some great laughs and he was always patient enough to hear the boring details of my research problems.

I thank all my current and former colleagues at the OPAC department for the good atmosphere and tea breaks. In particular, Claudine made for an excellent atmosphere in the E-corridor and she helped with many planning en technical problems (in the non-mathematical sense).

When I was young, my parents taught me the value of getting a good education. I thank them for supporting me all the way to a PhD degree. My siblings were very influential in shaping my ideas about the quest for truth and the role of academic research in this quest.

Doing a PhD can put quite some strain on family life and I thank my wife Heidi for supporting me in ways too numerous to mention. She performed a very thorough proofreading of the entire thesis and carefully checked my algebra when I was doing a long derivation and needed a second pair of eyes. My two sons were also instrumental in the moral support needed to complete a thesis. They make a short appearance in Chapter 5.

Finally, I thank my Father in heaven for his support in all things.

Contents

Chapter 1

Introduction

"The whole is more than the sum of its parts"

Aristotle

Interchangeable parts have revolutionized modern manufacturing. Before the industrial revolution, products were made one-by-one as a whole in workshops by craftsmen. After the industrial revolution and up until today, most products are assembled from interchangeable parts. Recent research identifies the French General Jean-Baptiste de Gribeauval as the principal originator of working with interchangeable parts (Hounshell, 1984). He introduced a system for constructing French artillery from interchangeable parts in 1765. His system became known as "le système Gribeauval". Thomas Jefferson[1] was introduced to le système Gribeauval in 1785 when he was in France as a diplomat, and visited the weapon workshop of Honeré Blanc who was a gunsmith implementing le système Gribeauval. In a letter to John Jay[2], Thomas Jefferson wrote about this visit:

> An improvement is made here in the construction of the musket which it may be interesting to Congress to know, should they at any time propose to procure any. It consists in the making every part of them so exactly alike that what belongs to any one may be used for any other musket in the magazine. The government here has examined and approved the

[1] Thomas Jefferson (1743-1826) is a founding father of the United States of America, its third president, and one of the principal authors of the declaration of independence.

[2] John Jay (1745-1829) is a founding father of the United States of America and the first Chief Justice of the United States

> method, and is establishing a large manufactory for this purpose. As yet the inventor[3] has only completed the lock of the musket on this plan. He will proceed immediately to have the barrel, stock and their parts executed in the same way. Supposing it might be useful to the U.S. I went to the workman. He presented me with the parts of 50 locks taken to pieces and arranged in compartments. I put several together myself taking pieces at hazard as they came to hand, and they fitted in the most perfect manner. The advantages of this, when arms need repair, are evident.[4]

Interchangeable parts enabled the industrial revolution in large measure because they enabled the division of labor (Smith, 1776) and therefore raised productivity. In the time of Thomas Jefferson however, producing muskets with interchangeable parts lowered productivity because it required resolving several metrological issues in order to make the parts truly interchangeable. Jefferson was aware of this; the advantage he saw, as is evident in his last sentence above, was in the repair and maintenance of the muskets *after* production on the battlefield. In fact, years later, Thomas Jefferson wrote to the secretary of war, Henry Knox[5], that Blanc's:

> ...method of forming the firearms appears to me so advantageous, when repairs become necessary, that I have thought it my duty not only to mention to you the progress of this artist[6], but to purchase and send you half a dozen of his officers fusils.[7]

The use of interchangeable parts and the division of labor continued to enable the mass production that has shaped modern society. It appears that the main reason to pursue the use of interchangeable parts, at least for Thomas Jefferson, was to facilitate maintenance operations. This was despite the fact that in the short run, producing with interchangeable parts was more expensive because the industrial metrology technology needed to make truly interchangeable parts had not yet been developed. This is striking: One of the most revolutionary ideas of modern manufacturing started out as a maintenance innovation!

This maintenance innovation also changed maintenance operations. Rather than performing maintenance or repair on equipment in its entirety, parts of equipment that require maintenance or repair are interchanged with *ready-for-use* spare parts.

[3]Thomas Jefferson here refers to Honoré Blanc

[4]Papers of Thomas Jefferson, 8:455, August 30 1785

[5] Henry Knox (1750-1806) was an officer, initially in the continental army during the American revolutionary war, and later in the United States army. Under the US presidency of George Washington, he was the secretary of war.

[6]Thomas Jefferson is again referring to Honoré Blanc

[7]Papers of Thomas Jefferson, 15:422, September 12 1789

After this, the equipment returns to serviceable condition immediately, while repair and maintenance is conducted on the replaced parts. This method for maintaining equipment greatly increases the availability of equipment. Since equipment often represents substantial financial investments (just think of aircraft, rolling stock, MRI-scanners and military equipment) achieving a high availability of equipment is important. Spare parts are essential in ensuring proper operation of this system of maintenance by replacing parts, and so they also affect daily services provided by equipment such as public transportation (rolling stock, aircraft) health care (MRI-scanners), and military operations (military equipment and weapon systems). The planning and control of spare parts for the support of maintenance operations is the topic of this thesis.

Maintenance spare parts planning and control also has a significant financial impact. Some impressive statistics that illustrate this are:

- In 2003, spare parts sales and services (mostly maintenance) accounted for 8% of the gross domestic product in the United States (AberdeenGroup, 2003).
- More recently, US bancorp estimated that the yearly expenditure in the US on spare parts amounts to 700 billion dollars which is 8% of the US gross domestic product (Jasper, 2006).
- A study by Deloitte (2006) among 120 large manufacturing companies in North America, Asia Pacific and Europe shows that service revenues represent more than 25% of total business.
- According to the same study by Deloitte (2006), aftermarket service and spare part sales account for 40% of profits for these 120 manufacturing companies.

Another indicator of the importance of spare parts and after sales services and maintenance is that original equipment manufacturers (OEMs) are increasingly realizing the potential, and are making a business model out of providing after sales services (Oliva and Kallenberg, 2003; Wise and Baumgartner, 1999). Nevertheless, in this thesis, we take the perspective of the owner of equipment that decides to keep maintenance and spare parts planning in house.

In the remainder of this introductory chapter, we discuss the industrial setting for which the models in this thesis have been developed. We start by giving a brief overview of maintenance operations and the supply chain of maintenance spare parts in §1.1 and §1.2. Here and throughout the thesis, we take the perspective of owner/maintainer of equipment rather than that of an OEM providing maintenance for its customers. We identify several issues that arise in the planning and control of maintenance spare parts and formulate the research objectives of this thesis in §1.3. We conclude this chapter by giving an outline of the thesis in §1.5.

1.1. Maintenance operations

Different from regular production operations, maintenance operations are not instigated by demand from an outside customer, but by the need for maintenance of equipment. To perform maintenance, typically several resources are needed, the most important of which are:

- a specialist, mechanic, engineer or other trained professional
- tools and equipment
- spare parts.

In §1.1.1, we discuss different maintenance strategies and how they instigate the need for maintenance operations (and therefore also the resources mentioned above). The planning difficulties that arise in maintenance operations are discussed in §1.1.2. The role of spare parts in maintenance operations is discussed in §1.1.3. By way of example, we conclude with a brief description of actual maintenance operations at NedTrain.

1.1.1 Maintenance strategies

For the purpose of describing maintenance operations, it is convenient to think of equipment as a collection of interrelated parts. Maintenance operations consist largely (but not solely) in replacing parts of equipment. Maintenance strategies determine when parts or equipment need to be replaced or maintained. Throughout this subsection, we focus on the decision to maintain/replace a part, but our discussion also applies to the decision to maintain/replace equipment. Figure 1.1[8] gives an overview of maintenance strategies. In this subsection, we follow Figure 1.1 in discussing different maintenance strategies.

Modificative maintenance concerns interchanging a part with a technically more advanced part in order to make the equipment perform better[9]. This form of maintenance is usually project based and non-recurring. The maintenance strategies that occur most often are preventive and breakdown corrective maintenance. Under a breakdown corrective maintenance strategy, a part is not replaced until it has failed, while under a preventive maintenance strategy, the aim is to replace parts before failure occurs. (Off course, this aim may not always be achieved: A part can break down before its replacement occurs.) Breakdown corrective maintenance is an

[8] Figure 1.1 was inspired by Figure 4.1 of Coetzee (1997), but has been significantly altered by the author.

[9] Sometimes maintenance is defined as any action that restores equipment to some previous state. Under this definition, modificative maintenance is an oxymoron.

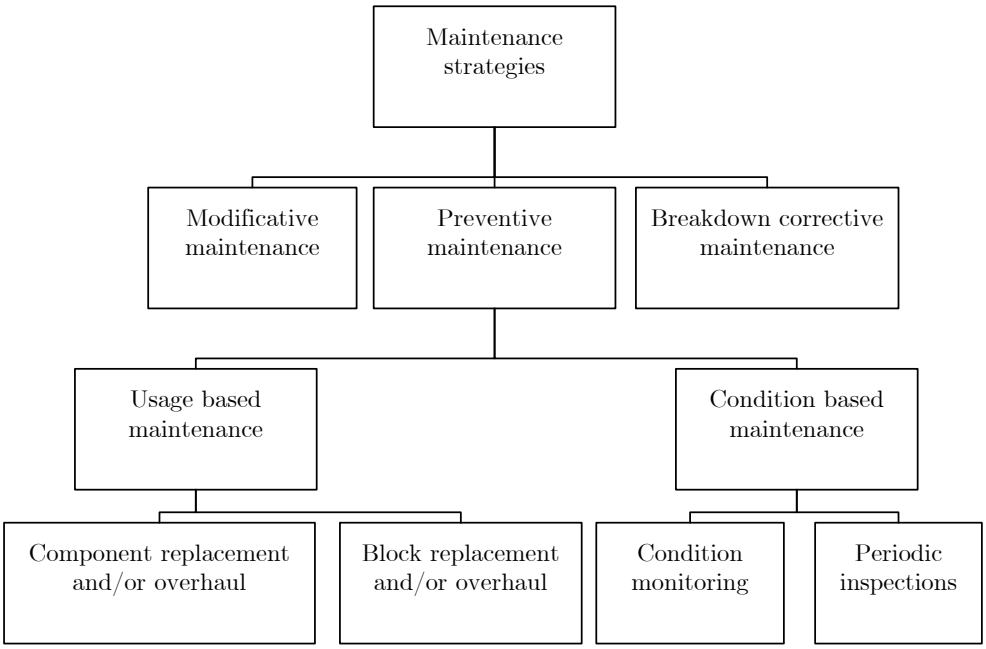

Figure 1.1 Maintenance strategies

attractive option for parts that do not wear, such as electronics. For parts that do wear, it can be beneficial to follow a preventive maintenance strategy.

Preventive maintenance strategies can be further divided into usage and condition based maintenance. Under usage based maintenance, the total usage of a part is measured and maintenance is conducted when a certain threshold level has been reached. The usage of parts can be measured in many ways depending on the nature of the equipment. Time in the field is perhaps the most common mean to measure usage. For vehicles (e.g., rolling stock), mileage is a common measure of usage. The number of on-off cycles is a measure of usage for equipment that is mainly loaded at the end or beginning of on-off cycles. For example, the number of landings is a measure of usage for the landing gear of an aircraft. Since the usage of equipment is usually scheduled, the moment that maintenance is performed can also be scheduled. If there is a large set-up cost associated with maintenance, it can be beneficial to interchange several parts simultaneously (Block replacement and/or overhaul). Otherwise, maintenance can be performed on a single component (Component replacement and/or overhaul).

In condition based maintenance, the actual condition of a part is gauged and maintenance is conducted based on this. The condition of a part can be measured

either periodically during inspections (Periodic inspections) or continuously through a sensor (Condition monitoring). How the condition of equipment is measured depends on the nature of equipment. Below are some examples of how the condition of equipment can be measured:

- The condition of ball-bearings can be measured via the amplitude of vibrations around the bearing (Elwany and Gebraeel, 2008).
- The condition of a metal part can be determined by visually inspecting the number and length of cracks.
- For metal systems with moving parts, the concentration of ferrous parts in the lubrication fluid is measured as an indication of the wear and need for lubrication.
- The condition of a car engine is monitored continuously while driving by the engine-oil temperature gauge.

The need for maintenance can be ascertained periodically during an inspection or at any time in case of condition monitoring.

Which types of maintenance are prevalent for a given piece of equipment depend very much on the technical nature of the equipment involved. For electronics and high-tech equipment, breakdown corrective maintenance is prevalent. For aircraft, rolling stock and other heavy machinery with moving parts, the prevalent maintenance strategies are preventive (both usage and condition based).

1.1.2 Uncertainty in maintenance operations

Maintenance operations are subject to considerable uncertainty. There is uncertainty both with respect to timing (When will maintenance/replacement be needed?) and content (What parts need maintenance/replacement?). The different maintenance strategies discussed in the previous subsection are organized according to these two uncertainty dimensions in Table 1.1[10]

Usage based and modificative maintenance can be planned for ahead of time, whereas breakdown corrective maintenance cannot be planned for at all. As a consequence of this, the resources needed for usage based and modificative maintenance can be utilized more fully than resources needed for breakdown corrective maintenance.

[10]Table 1.1 has been inspired by the maintenance box of Stoneham (1998) but has been altered significantly by the author.

Table 1.1 Maintenance strategies organized by timing and content uncertainty.

		Timing	
		known	unknown
Content	known	Usage based or modificative maintenance	Condition based maintenance (Condition monitoring)
	unknown	Condition based maintenance (Periodic inspections)	Breakdown corrective maintenance

Condition based maintenance is a hybrid form, in which some but not all uncertainty is taken away relative to breakdown corrective maintenance. Periodic inspections can be planned, and if they lead to maintenance, you know when the maintenance needs to be conducted (right after the inspection). However, the content of the maintenance depends on what is found during the inspection. Under condition monitoring, sensors provide realtime information about the degradation of equipment. The parts that need replacement can then be inferred from the sensor signal. However, degradation usually remains an uncertain process, so that the exact time that maintenance is needed remains unknown.

Remark 1.1 Sometimes the distinction between preventive and corrective maintenance is interpreted as being synonymous to planned and unplanned maintenance. This oversimplification only captures the upper left and lower right boxes of Table 1.1. Condition based maintenance is a hybrid form between planned and unplanned maintenance that deserves separate attention. ◇

1.1.3 Spare parts in maintenance operations

In this thesis, we distinguish three different types of maintenance spare parts:

- Rotables - These are items that constitute a sufficiently large subsystem of the original equipment to warrant a separate usage based maintenance strategy. Rotables are individually tracked and traced so that the correct usage can be ascribed to each rotable individually. Usually, there are dedicated resources for the maintenance and overhaul of rotables. Examples include aircraft engines, rolling stock bogies (see Figure 1.2a), and elaborate weapon or radar systems on frigates.
- Repairables - These are items that are repaired after replacement after which they are *ready-for-use* (RFU) again. Contrary to rotables, repairables do

(a) A bogie is an example of a rotable.

(b) A compressor is an example of a repairable.

(c) A gasket is an example of a consumable.

Figure 1.2 Examples of different types of spare parts.

not have their own usage based maintenance strategy, and are not usually individually tracked and traced. A repair shop handles the repair of many different types of repairables. Examples of repairables include compressors (see Figure 1.2b) and pumps.

- Consumables - These are items that are discarded after replacement and bought new from a supplier. Generally these are relatively cheap items such as gaskets (see Figure 1.2c).

These different part types generally are also connected to different maintenance strategies as shown in Table 1.2. Demand for spare parts inherits the uncertainty characteristics of the type of maintenance for which they are used; see Table 1.1. For example, there is almost no demand uncertainty for rotables, while demand uncertainty for consumables subject to breakdown corrective maintenance is high.

Table 1.2 The role of different part types in maintenance operations.

	Type of spare part		
Maintenance strategy	Rotable	Repairable	Consumable
Usage based	x		
Condtion based		x	x
Modificative		x	
Breakdown corrective		x	x

As a note on terminology, we mention that spare parts used to maintain equipment are called line replaceable units (LRUs). (LRUs can be either rotable, repairable or consumable.) We will use the abbreviations LRU and RFU extensively in §1.2.

1.1.4 Maintenance operations at NedTrain

As indicated at the beginning of §1.1, we conclude this section with a brief description of maintenance operations at NedTrain. NedTrain is a division of the Dutch railways (NS, Nederlandse Spoorwegen) that is responsible for the servicing and maintenance of all rolling stock of the NS. NedTrain operates four large maintenance depots throughout the Netherlands[11]. The fleet of trains they maintain has around 2800 coaches. Each rolling stock unit (which consists of several coaches) visits one of the maintenance depots approximately every three months. Such a visit lasts one to several days. Most of the maintenance at NedTrain is condition based. During a visit to the maintenance depot, inspections are performed on the rolling stock and maintenance is done according to the outcome of these inspections.

Certain new rolling stock units have sensors on board that enable condition monitoring. Currently, the output of these sensors is primarily used to quickly identify the culprit after breakdown. It is expected that in time, condition monitoring can be used to assess the condition of rolling stock before entrance into the depot, and can even influence the moment of depot entrance.

NedTrain applies usage based maintenance for several rotables and has dedicated repair and overhaul facilities for these rotables. The rotables are replaced in the maintenance depot and receive thorough revision and overhaul in a specialized shop.

1.2. Spare part supply chains

Figure 1.3 gives an overview of a typical spare part supply chain. Demand for LRUs occurs at one or more maintenance depots where equipment is maintained. There are stock points incident to the maintenance depots where ready-for-use spare parts are kept. These stock points are supplied from a central stock point which in turn is supplied by external suppliers (in case of consumables), and internal and external spare part repair shops (for repairables and rotables). There is a return flow of repairable and rotable LRUs that require either repair or maintenance and overhaul. Stock is also kept for modificative maintenance. This stock point may be physically located at one of the other RFU-LRU stock points, but it is typically controlled separately.

If the shipment time of a part from one RFU-LRU stock point to another is small compared to the replenishment lead time of the central warehouse, the entire network

[11] NedTrain also operates several other locations called service depots or technical centers where daily services, cleaning and small repairs are conducted. Since these maintenance operations are small compared to the maintenance in maintenance depots, we do not discuss them in detail here.

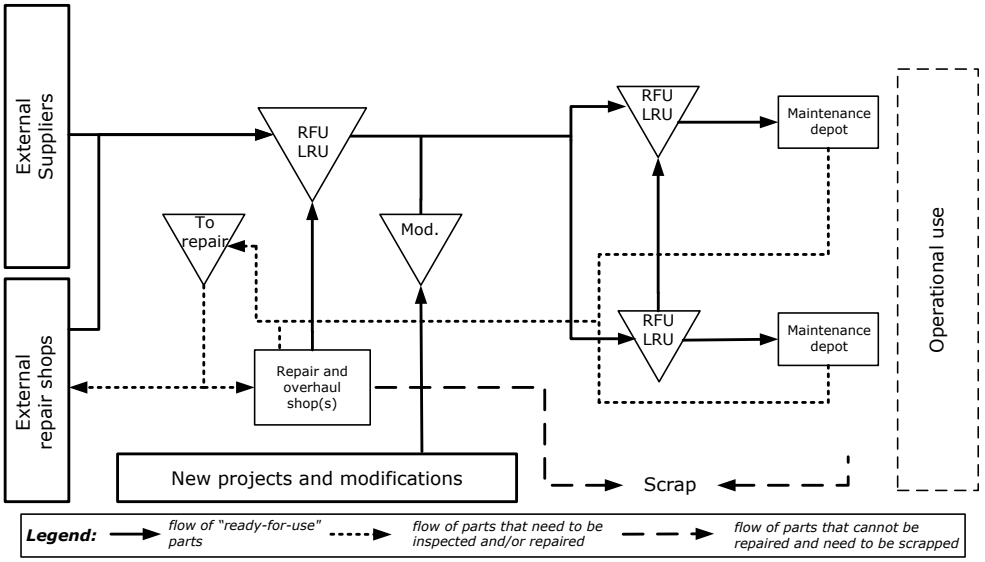

Figure 1.3 Typical example of a maintenance spare parts supply chain.

of RFU-LRU stock points can be considered as one big virtual stock point. In the remainder of this section, we discuss repair and overhaul shops (§1.2.1), performance measures (§1.2.2) and conclude with a brief description of an actual spare part supply chain as found at NedTrain.

1.2.1 Repair and overhaul shops

A crucial role in the spare parts supply chain is played by the repair and overhaul shops. Since the equipment is mostly maintained by interchanging parts, the actual repair and maintenance occurs mostly at the part level and is done in repair and overhaul shops. Performing the repair of parts is usually not a very standardized process because different parts of the same type may fail for widely varying reasons. Furthermore, repair shops often perform repairs for many different repairable types. Consequently, repair and overhaul of spare parts is usually a labor intensive process (Paz and Leigh, 1994). Shops that deal with repairables are usually referred to as repair shops, whereas the shops deal with rotables are usually referred to as overhaul shops.

1.2.2 Performance measures

The whole spare parts supply chain exists solely to make maintenance operations run smoothly, such that equipment is down for maintenance no longer than necessary. The performance of a spare parts supply chain with respect to maintenance operations is commonly measured by several performance measures such as:

- The total expected number of backorders
- The waiting time for an arbitrary request for a part
- The fraction of part requests that can be met immediately (fill-rate)

For a more thorough discussion of performance measures in spare part supply chains and their relations to each other, we refer the reader to Vliegen (2009) and Dinesh Kumar et al. (2000). In this thesis, we focus on the total expected number of backorders. If we neglect the possibility that a single piece of equipment is waiting for more than one part on backorder, the expected number of backorders corresponds (approximately) to the number of pieces of equipment that is down waiting for a spare part to become available. Note that the expected number of backorders for some specific spare part is not directly of interest; rather we are interested in the total number of backorders for types of spare parts together.

The obvious other performance measure of a spare parts supply chain are the costs it incurs. These costs depend on many things but the two main cost drivers are the number of spare parts and the capacity available in repair shops.

1.2.3 Spare part supply chain at NedTrain

NedTrain has four stock points incident to their maintenance depots as well as a central stock point. They have two repair shops, one of which mainly repairs and overhauls rotables while the other mainly repairs repairables. External repair shops are contracted for certain specialized types of repair. The repair lead time of the repair shops is in the order of magnitude of several weeks. The replenishment lead time of different consumables varies widely from several hours for very cheap parts to several months for expensive consumables.

Shipment times from one RFU-LRU stock point to another are several hours up to a day. This is very short compared to the replenishment lead times at the central stock point (typically several weeks) and compared to the time equipment is in the maintenance depot (one up to several days). Furthermore, emergency lateral transshipments between stock points occur when needed and take no more than a few

hours. Thus, it is reasonable to consider the entire network of RFU-LRU stock points as one virtual RFU-LRU stock point.

The repair shop uses priority rules to schedule the repair of parts. These priority rules are designed to keep the stock of each repairable item above agreed minimum levels. These minimum levels are set such that priority is given to the repair of parts for which there is little stock or for which a demand surge is expected to occur.

1.3. Research objectives

In the previous two subsections, we have briefly sketched the environment for which the models in this thesis apply. Now we discuss the research objectives of this thesis. At this stage of the thesis, we do not position our research objectives relative to existing literature. This is done in more detail in each chapter individually.

1.3.1 Framework

Planning and controlling maintenance spare parts entails many different aspects. Models in literature typically focus on one or a few of those aspects, as will most models in this thesis. But before we focus on a few specific aspects, we would like to gain a broader understanding of all the aspects and their interrelationships on a qualitative level. This broad understanding should be helpful see the quantitative models in this thesis within a broader perspective. It should also be helpful for professionals in practice. We formulate the following research objective:

Research objective 1 Develop a framework for the planning and control of a spare part supply chain in organizations that own and maintain equipment. This framework should outline all relevant decisions that are made in such a supply chain and explain how they relate to each other.

1.3.2 Rotables, usage based maintenance and efficient utilization of resources

Rotables are spare parts, but because they have a usage based maintenance strategy, demand for these items is predictable. After rotables are replaced, they are overhauled in a repair shop. The resources needed for overhauling are expensive and so it is important to make effective use of these resources. Therefore, we formulate the following research objectives:

Research objective 2 Develop a planning algorithm that makes efficient use of the resources needed for rotable replacement and overhaul. This algorithm should exploit the fact that demand for rotables is predictable.

Research objective 3 Investigate the value of using the predictability of demand for rotables in making efficient use of resources.

1.3.3 Repairables, condition based maintenance, and repair lead time flexibility

In §1.1.2, we observed that condition based maintenance is greatly affected by uncertainty, but much less so than for breakdown corrective maintenance. This should naturally translate into a better understanding of the demand process of repairables. Empirical evidence (e.g Slay and Sherbrooke, 1988) suggests that demand is not stationary in these cases. In practice, this non-stationarity is buffered by both inventory and smart scheduling in the repair shop. Smart scheduling in the repair shop leads to dynamic lead time flexibility: Repairable parts for which demand is momentarily high and RFU inventory is low are temporarily endowed with shorter lead times, while other parts with sufficient on-hand inventory and momentarily low demand will temporarily experience long lead times. The research objectives below are aimed at understanding how lead time flexibility, repairable inventory and non-stationary demand arising from condition based maintenance interact, and what the value is of exploiting lead time flexibility and knowledge from condition based maintenance about the non-stationarity of demand.

Research objective 4 Develop a model of repair lead time flexibility and non-stationary demand due to condition based maintenance for a single-item and investigate how information regarding demand non-stationarity from condition based maintenance can be used to leverage repair lead time flexibility.

Research objective 5 Develop a model that can assess the interplay between repairable inventory and lead time flexibility in buffering demand uncertainty and non-stationarity.

Research objective 6 Investigate the value of explicitly modeling lead time flexibility and demand information arising from condition based maintenance.

After we reach all the objectives above, the next step is to develop a model that can aid decision making. In particular the initial supply decision is important. Along with buying equipment, there usually is the possibility to purchase repairable spare parts

at a reasonable price. This decision should reflect the fact that due to condition based maintenance, demand will not be stationary and repair lead times can be influenced to deal with these demand fluctuations.

Research objective 7 Develop a tractable multi-item optimization algorithm that supports the initial supply decision and incorporates lead time flexibility, non-stationary demand arising from condition based maintenance and performance objectives on fleet level.

1.3.4 Consumables, emergency procedures

Consumables that wear out quickly are usually cheap compared to repairables. Because they are replaced regularly, demand for such consumables is stationary. However, when a consumable is out of stock, it is common practice to fulfill demand for this consumable by an emergency procedure. For example, the consumable may be picked up by a mechanic at a local hardware store. This emergency procedure is expensive and causes the original stock point not to see this demand, which is analogous to a lost sale in a retail environment and can be modeled the same way.

Unfortunately, lost sales inventory problems are known to be difficult, both to optimize and to estimate the performance of. The optimal replenishment policy for such inventory systems is not well understood or easy to compute. The periodic review base-stock policy is commonly used in practice, but even this policy is difficult to optimize and to evaluate the performance of. Our objective is therefore the following.

Research objective 8 Develop an algorithm for the optimization of the base-stock level in a periodic review lost sales inventory system that is fast and provides accurate estimates of performance measures.

1.4. Contributions of the thesis

In §1.1.3, we described three types of spare parts: rotables, repairables, and consumables. The contributions we make in this thesis are best organized by these 3 spare part types.

1.4.1 Rotable overhaul planning

Chapter 3 studies the scheduled usage based maintenance of rotable parts. Usage based maintenance policies stipulate that a rotable should not be used any longer

that the maximum inter-overhaul time (MIOT). Traditional approaches to scheduling usage based maintenance focus on postponing overhaul as long as possible to take advantage of the technical life of the rotable. Models accomplish this by planning only one overhaul instant into the future and artificially penalizing early overhaul of a rotable. These penalty costs are really just a proxy for minimizing the amount of maintenance conducted over the entire lifetime of a piece of equipment. The underlying assumption is that this will also minimize the costs of materials and required capacity in the overhaul workshop.

Our approach in Chapter 3 is more direct because we consider the costs of material and overhaul capacity over the entire lifetime of the equipment directly, rather than indirectly via an artificial penalty parameter. This approach can exploit opportunities for cost savings that the traditional approach cannot. We illustrate this with the following example from NedTrain.

The typical lifetime of a rolling stock unit is 30 years. Bogies are important rotables in a train, with MIOTs that range from 4 to 10 years. Suppose the MIOT of two types of bogies is 7 years, and both types of bogies belong to the same type of train. Then, if replacements are planned to occur just in time, bogie replacements occur 4 times during the life cycle of this train type, namely in years 7, 14, 21, and 28. This plan can be modified *without* changing either the amount of material needed over the lifetime of the train, or the overhaul capacity: Overhaul rotables in years 6, 12, 19, and 25. By sticking to the original plan for one type of rotable and changing to the other plan for the other type, we can reduce peak overhaul capacity needed, and, therefore, we can reduce overhaul capacity levels and costs.

In effect, we are not, and should not, be concerned with minimizing the amount of useful lifetime on rotables that is wasted in the short run. Rather, we should minimize the cost of overhaul that rotables incur over the entire lifetime of the equipment they serve, which is finite.

The difficulty in formulating the planning problem above is the propagation of overhaul deadlines over a long planning horizon. (Recall that traditional models only plan one overhaul into the future.) With some auxiliary variables, we formulate this planning problem as a mixed integer linear programming problem (MIP). We show that this problem is strongly $\mathcal{NP}$-hard, but also provide computational evidence that our MIP formulation can be used to solve real life instances. We also provide computational evidence that the linear programming relaxation of the the MIP formulation is quite tight and can be used for sensitivity analyses.

1.4.2 Repairable stocking and expediting under fluctuating demand

Repairable spare parts are expensive and in many practical situations, it is not possible to buy new repairables at will. The best time for companies to buy repairables is at the same time as when the original equipment is purchased. Nevertheless, demand for repairable items typically fluctuates over time, reflecting the fluctuating need for maintenance over time. Companies anticipate these demand fluctuations by leveraging the possibility of expediting the repair of defective parts, rather than buying new parts. Expedited repairs have a shorter lead time but incur additional costs per repair job.

Chapter 4 studies the situation described above and supports two decisions at the tactical and operational level respectively:

1. How many repairable spare parts should we buy? (tactical)
2. When should we request that the repair of a part is expedited? (operational)

We study this decision problem via a stochastic inventory model for repairable items. In this model, a defective item is replaced with a ready-for-use item and sent to a repair shop immediately after the defect occurs. At this point in time, the inventory manager is faced with the decision to either expedite or not expedite the repair of the part. This expediting decision is informed by knowledge about the fluctuation of demand intensity over time. The fluctuation of demand over time is modeled by a Markov modulated Poisson process. The state of the Markov chain that drives the demand process can be observed directly and is used to inform the expediting decision. We assume that repairable item inventory is replenished by a lot-for-lot policy (as is common in practice). We model the expedited lead time as being deterministic and the regular lead time as being the convolution of the expedited lead time and several exponential phases, the passing of which is monitored. Many lead time distributions can be modeled quite closely by this device and in particular deterministic lead times can be approximated arbitrarily closely by letting the number of exponential phases approach infinity.

The main contributions of chapter 4 are as follows. For the described setting, we characterize the optimal repair expediting policy for the infinite horizon average and discounted cost criteria by formulating the problem as a Markov decision process. We find that the optimal policy may take two forms. The first form is simply to never expedite repair. The second form is a state dependent threshold policy, where the threshold depends on both the state of the modulating chain of demand and the pipeline of repair orders. We also provide monotonicity results for the threshold as a function of the pipeline of repair orders. We give closed-form conditions that determine which of the two forms is optimal. In analyzing the optimal policy, we

also confirm a conjecture of Song and Zipkin (2009) that the expediting policy they propose is optimal for some special cases.

Secondly, we show how to optimally solve the joint problem of determining the turn-around stock and the expediting policy.

Thirdly, we propose a heuristic that is computationally efficient, and is shown to perform well compared to the optimal solution. In this heuristic, we replace the optimal expediting policy with a parameterized threshold policy that shares important monotony properties with the optimal expediting policy. The thresholds depend on the available knowledge about the fluctuation of demand. Borrowing the terminology of Zipkin (2000), we call this policy the world driven threshold (WDT) policy. In a numerical study involving a large test bed, this heuristic has an average and maximum optimality gap of 0.15% and 0.76% respectively.

Finally, we investigate the value of anticipating demand fluctuations by comparing optimal joint stocking and expediting policy optimization against naive heuristics that do not explicitly model demand fluctuations, or that separate the stocking and expediting policy decisions. These naive heuristics have optimality gaps of 11.85% on average and range up to 63.67% in our numerical work. The comparison with these naive heuristics show that

1. There is great value in leveraging knowledge about demand fluctuations, in making repair expediting decisions.
2. Fluctuations of demand and the possibility to anticipate these through expediting repairs should be considered explicitly in sizing the turn-around stock and can lead to substantial savings.

In Chapter 5, we extend the model of Chapter 4 to a multi-item multi-fleet multi-repair shop setting. The scheduling of repair jobs in the repair shop has long been known to have a significant effect on the inventory investment required to meet several common service levels. (Hausman and Scudder, 1982; Tiemessen and Van Houtum, 2012, e.g.). Optimal scheduling rules for capacitated repair shops are quite intractable and even the evaluation of simple priority rules suffers heavily from the curse of dimensionality. For this reason, simulation optimization with local search are the techniques most commonly used to determine good repair scheduling and repairable stocking policies.

Our model also considers the situation where scheduling in the repair shop can affect the repair lead time of parts, but we refrain from explicitly modeling the dynamics that occur on the shop floor. We assume that it is possible to expedite the repair of a limited number of repair jobs per time unit on average. This allows us to model the essential characteristics of smart scheduling policies, namely that the repair lead time can be shortened for parts that are in short supply and lengthened for parts that are in

ample supply. The merit of the model in chapter 5 is that it can do this in a tractable manner. Even so, our final model is non-linear non-convex integer programming problem. We show how to find lower bounds for this problem via a column generations algorithm in which the pricing problem is exactly the problem studied in Chapter 4. We also show how to obtain a good feasible solution within reasonable time using binary programming techniques. In extensive numerical experiments, the feasible solution we found had an optimality gap of 0.67% on average and 6.76% at most. We also quantify the effect of considering repair shop flexibility through expediting compared to models in which stocking decisions are based on a single mean lead time. Explicitly considering these flexible lead times through expediting leads to an average reduction in repairable spare parts investment of 25% compared to the approach based on a single lead time for a large test-bed.

1.4.3 Consumable stocking with emergency shipments

Chapter 6 studies base-stock policies for consumables that are reviewed periodically. When the stock for consumables is depleted, it is a common procedure to use an emergency supply source to replenish the part almost instantaneously so that maintenance is not halted for lack of a part. All items that are replenished by the emergency procedure are lost to the normal mode of replenishment. This problem is mathematically equivalent to the classical lost sales inventory problem that has been studied by Karlin and Scarf (1958), Janakiraman et al. (2007), Zipkin (2008b), Zipkin (2008a), Levi et al. (2008), and Huh et al. (2009b). This system consists of a periodically reviewed stock point which faces stochastic i.i.d. demand. When demand in a period exceeds the on hand inventory, the excess is lost. Replenishment orders arrive after a lead time τ. At the end of each period, costs for lost sales and holding inventory are charged. For such systems, we are interested in minimizing the long run average cost per period.

The structure of the optimal policy for lost sales inventory systems with a positive replenishment lead time is still not completely understood, and the computation of optimal policies suffers from the curse of dimensionality as the state space is τ-dimensional. Huh et al. (2009b) show that base-stock policies are asymptotically optimal as the lost sales penalty costs approach infinity. However, computing the best base-stock policy for a lost-sales inventory problem efficiently remains a challenge. Huh et al. (2009a), p. 398, observe that: "Although base-stock policies have been shown to perform reasonably well in lost sales systems, finding the best base-stock policy, in general, cannot be accomplished analytically and involves simulation optimization techniques". Although the burden of optimization is alleviated by the fact that the average cost under a base-stock policy is convex in the base-stock level (Downs et al., 2001; Janakiraman and Roundy, 2004), evaluating the performance of

any given base-stock policy requires either value iteration or simulation.

Chapter 6 provides an efficient method to compute near optimal base-stock levels for lost sales inventory models as well as accurate approximations for the costs of base-stock policies. This method is based on a different view of the dynamics of a lost sales inventory system, inspired by a relation to the dual sourcing inventory system. This relation has been studied by Sheopuri et al. (2010), and allows us to use ideas similar to those of Arts et al. (2011) for dual-sourcing inventory systems in the context of lost sales inventory systems. Somewhat counter-intuitively, our approach involves moving from a τ-dimensional state space description to a $(\tau + 1)$-dimensional state space description, where τ is the order replenishment lead time. This $(\tau + 1)$-dimensional state space is the pipeline of all outstanding orders, but not the on-hand inventory. The next key idea to this approach is to aggregate this pipeline of outstanding orders into a single state variable. This is essential to lending tractability as the size of the original state space grows exponentially in both the lead time *and* the base-stock level. By contrast, the aggregated state space grows linearly in the base-stock level only.

From the distribution of this single aggregated state variable, all relevant performance measures can be computed. The distribution of this single state variable can be studied via a Markov chain. For the transition probabilities of this Markov chain, we derive limiting results and show that for the most commonly used demand distributions, the rate of convergence for these limits is at least exponential. We also show that these limiting results satisfy a type of flow conservation property. This flow conservation property relates the average size of an order entering or leaving the pipeline to the total number of items in the pipeline. Based on these results, evaluating a single base-stock policy approximately is as easy as solving $S + 1$ linear equations, where S is the base-stock level. Numerical experiments indicate that this approach yields excellent results. Across a test bed that is an extension of the test beds considered by Huh et al. (2009b) and Zipkin (2008a), we find that our approach has cost differences with the best base-stock policy of at most 1.30% and 0.01% on average.

1.5. Outline of the thesis

Table 1.3 gives an overview of the thesis by the research objectives as stated in §1.3 and main methodology. Table 1.4 gives an overview of the material in this thesis based on spare part type and maintenance type. Chapters 2, 3 and 5 are intended to be accessible to a wide audience of both practitioners and academics interested in maintenance and spare parts planning and control. Chapters 4 and 6 contain more

specialized technical results that are mostly of interest for researchers in inventory theory and stochastic operations management. The chapters have been set up such that they can be read independently. A slight exception to this is chapter 5. For the detailed analysis of the model in chapter 5, we refer to some results in chapter 4. However, chapter 5 has been written such that the analysis section can be skipped by readers that are not interested in the mathematical details.

The work in chapter 2 is based on Driessen et al. (2010) and chapter 3 is based on Arts and Flapper (2013).

Table 1.3 Navigating the thesis by research objective and methodology.

	Research objective									
Chapter	1	2	3	4	5	6	7	8	9	Main methodology
2	x	x								Conceptual framework
3			x	x						Mixed integer programming
4					x	x	x			Markov Decision Process
5						x	x	x		Column generation
6									x	Asymptotics

Table 1.4 Navigating the thesis by spare part and maintenance type.

	Spare part type			Maintenance strategy		
Chapter	Rotables	Repairables	Consumables	Breakdown corrective	Usage based	Condition based
2		x	x	x		x
3	x				x	
4		x		x		x
5		x		x		x
6			x	x		x

Chapter 2

Maintenance spare parts planning framework

"Every theory is a self-fulfilling prophecy that orders experience into the framework it provides."

Ruth Hubbard

2.1. Introduction

Many industries depend on the availability of high-value capital assets to provide their services or to manufacture their products. Companies in these industries use capital assets in their primary processes and hence downtime can among others result in (i) lost revenues (e.g., standstill of machines in a production environment), (ii) customer dissatisfaction and possible associated claims (e.g., for airlines and public transportation) or (iii) public safety hazard (e.g., military settings and power plants). Usually the consequences of downtime are very costly.

A substantial group of companies in these industries both use and maintain their own high-value capital assets. Examples include airlines, public transportation and military organizations. Within these companies, a *Maintenance Organization* (MO) is responsible for maintaining the capital assets. Besides maintenance activities, supply and planning of resources, such as technicians, tools and spare parts, are required. A *Maintenance Logistics Organization* (MLO) is responsible for matching the supply

and demand of the spare parts required to conduct maintenance.

Because the capital assets are essential to the operational processes of the companies involved, downtime of the assets needs to be minimized. Downtime of a system is usually divided into (i) diagnosis and maintenance time; (ii) maintenance delay caused by unavailability of the required resources for diagnosis and maintenance. A high availability of spare parts is important as it influences the maintenance delay. In this chapter, we focus on the responsibility of a MLO to minimize maintenance delay due to unavailability of required spare parts.

Our main contribution is the development of a hierarchical framework for MLOs as described above. This framework outlines the decisions that need to be made to effectively control a spare parts supply chain. It also describes the interactions and (hierarchical) relations between these decisions and provides an outline of how these decisions can be decomposed. As such, the framework it is a type a taxonomy of different decision functions and their interrelations; see Figure 2.3 for a quick graphical overview of the framework. We also embed the framework in existing literature by reviewing results and models that can be used to support most of the decisions in the framework. In performing this review, we also highlight a few areas that deserve additional research.

This framework also serves as a useful starting point in making specific designs of maintenance spare part planning and control systems. For organizations with an existing design, the framework has a mirror function. That is, it can be used to compare the current design of the spare parts planning and control at a given company to our framework. Such comparative studies, based on the framework in this chapter, have successfully been conducted at different companies from different industries:

- Railway industry (Driessen and Arts, 2011, NedTrain)
- Airforce (Driessen, 2011, Royal Dutch Airforce)
- Aviation industry (Driessen, 2012a, Royal Dutch Airlines (KLM) Engineering and Maintenance)
- Army (Driessen, 2012b, Royal Dutch Army)
- Navy (Driessen, 2012c, Royal Dutch Navy)
- Port industry (Driessen, 2013, Europe Container Terminals)

Practitioners in all these case studies found the framework particularly useful to to increase the efficiency, consistency and sustainability of decisions on how to plan and control the spare parts supply chain.

To decompose decisions in a hierarchical framework is a well established approach in operations management. Initial models consider especially the production environment (Hax and Meal, 1975; Bitran and Hax, 1977; Hax, 1978; Bitran et al., 1981, 1982) and were motivated by the fact that it is computationally infeasible to solve one single all encompassing model. Later it was recognized that the hierarchy in such models is also useful because

(i) In reality the power to make decisions is distributed over several managers or agents;

(ii) The information available for different decisions has varying levels of detail (Dempster et al., 1981; Meal, 1984; Schneeweiss and Schröder, 1992; Schneeweiss, 1998, 2003; Schneeweiss and Zimmer, 2004).

For the production environment, hierarchical frameworks are now part of standard textbooks (Silver et al., 1998; Hopp and Spearman, 2001). Other successful applications include traffic control (Head et al., 1992) and supply chain management (Schneeweiss and Zimmer, 2004; Ivanov, 2010). Practitioners within the general discipline of supply chain management have also created a framework to facilitate their work (Council, 2010).

For spare parts specifically, such frameworks/guidelines and standards exist within the United States department of defense (US-DoD); see MIL-HDBK-965 (1996); MIL-PRF-49506 (1996); MIL-STD-1390D (1993); MIL-STD-3018 (2011). The standards and handbooks of the US-DoD focus primarily on setting up contracts with suppliers to make sure that there is an acquisition contract for each relevant type of spare part. These standards also adres quality and reliability standard for spare parts as well as standardization guidelines. Our framework takes a broader perspective by also considering the logistical control of the spare parts supply chain. We adres such issues as inventory control policies , repair shop control (for repairable spare parts), and spare part demand forecasting.

This chapter is organized as follows. §2.2 describes the environment we investigate and the positioning of MLOs. §2.3 presents the framework and describes the decisions in the framework. §2.4 provides relevant references for each part of the decision framework to aid in decision making and discusses open research topics. In §2.5, we give concluding remarks.

2.2. Characterization of the environment

In the primary processes of the companies we consider, a substantial set of capital assets (installed base) is used for multiple purposes. Because of strategic decisions, new systems phase in and other systems phase out of the installed base. Maintaining this set of assets is an important task because downtime of assets immediately affects the primary processes. A capital asset is (partially) operational in case it is available for (a part of) all its assigned purposes and a capital asset is down whenever it is in maintenance or waiting for maintenance to be conducted. Maintenance conducted within the constraints of the maintenance policy/concept can fall in any of the four categories of the maintenance box in Table 1.1.

To reduce the time an asset spends in maintenance, it is common practice to maintain parts of the asset rather than the asset itself. When an asset is maintained, parts that require repair are taken out and replaced by *ready-for-use* (RFU) parts. Spare parts used at the first level maintenance are also called *Line Replaceable Units* (LRUs) (Muckstadt, 1973, 2005). The decision to designate a part as a LRU lies with the maintenance organization. LRUs that are taken out are either scrapped or sent to a repair shop for repair. Repaired parts are sent back to a ready-for-use LRU stocking location where they can be used again to replace a part. This principle is called 'repair-by-replacement' (Muckstadt, 2005) and makes the control of the spare parts supply chain a paramount task for the MLO.

MLOs try to find the optimal balance between spare parts availability, working capital and operational costs, within their span of control. Several tasks need to be conducted and decisions need to be taken in order to achieve the desired spare parts availability, possibly under constraints of working capital and/or operational costs.

In this section, an outline is given of the environment in which MLOs operate. First, we characterize the process of maintaining the capital assets, second we discuss the spare parts supply chain and we end with the characterization of spare parts demand.

2.2.1 Characterization of system maintenance

The MOs we consider maintain a fleet of high-value capital assets. The installed base is sufficiently large to generate a reasonably constant demand for maintenance activities. Examples of such installed bases include fleets of airplanes, trains, weapon systems, or manufacturing equipment in a large manufacturing facility. Maintenance on a capital asset is conducted according to a maintenance policy/program, or a modification plan. We distinguish three types of maintenance from the engineering perspective:

- Preventive maintenance: maintenance that is conducted in order to prevent failure. Usually this maintenance is planned some time in advance and has to be conducted within a registered time frame during which the asset is in non-operating condition.
- Corrective maintenance: maintenance that is conducted after a failure has occurred. Corrective maintenance can be partially planned when it involves a non-critical part whose maintenance can be delayed.
- Modificative maintenance: maintenance conducted to improve the performance of the capital asset. This maintenance can usually be delayed until all resources are available.

With regard to the logistics of maintenance, we distinguish four types that do not always map directly to the three engineering types of maintenance. These four types are best understood by adapting the maintenance box in Table 1.1 to the present context; see Table 2.1.

Table 2.1 Different maintenance types organized by timing and content uncertainty.

		Timing	
		known	unknown
Content	known	Preplanned modificative maintenance	Condition based maintenance (condition monitoring)
	unknown	Condition based maintenance (periodic inspections)	Breakdown maintenance

Modificative maintenance can be planned carefully to take away any uncertainty with respect to the maintenance work and maps directly to modificative maintenance in the engineering sense. The other types of maintenance in Table 2.1 do not map so neatly into the engineering types of maintenance. While breakdown maintenance is always corrective maintenance, condition based maintenance can be either preventive or corrective. From a logistics point of view, breakdown maintenance is unpredictable, while condition based maintenance can be predicted to varying degrees.

The MO has to plan all the types of maintenance shown in Table 2.1. Figure 2.1 presents a hierarchical planning framework for maintenance of capital assets. The figure is to be read top-down. Work orders generate demand for LRUs and other resources (technicians, tools, equipment) needed to conduct the maintenance. A work order is released as soon as all resources and all required LRUs to start the work order are available. Unreleased work orders are queued, until they are released. The MLO

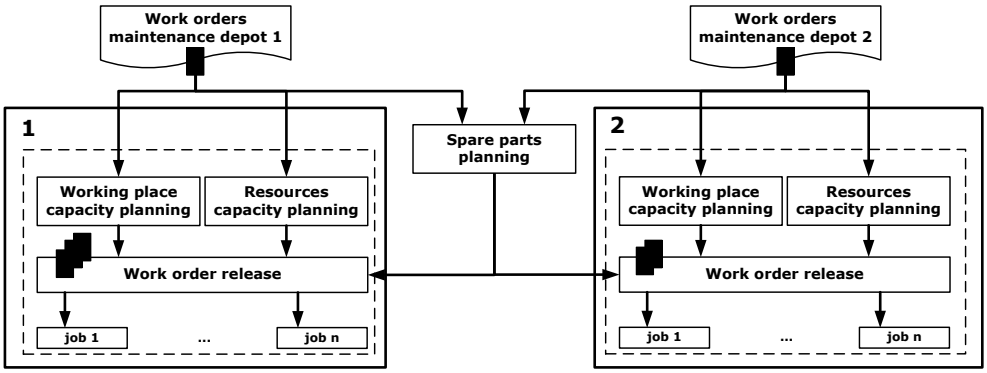

Figure 2.1 Hierarchical planning framework for maintenance of high-value capital assets.

is responsible for the availability of LRUs needed to conduct system maintenance, the MO is responsible for all other resources.

2.2.2 Maintenance spare parts supply chain overview

We consider organizations in which the supply chain already exists, i.e., location and size of warehouses are predetermined. The spare parts supply chain is in general a multi-echelon system. We distinguish two types of spare parts:

1. Repairable parts: parts that are repaired rather than procured, i.e. parts that are technically and economically repairable. After repair the part becomes ready-for-use again.

2. Non-repairable parts or consumables: parts which are scrapped after replacement.

In §1.1.3, we discussed rotables[1] as a third type of spare part. The control of rotables is out of scope for the present chapter, but their planning and control is addressed in Chapter 3. Consumable LRUs need to be replenished from outside suppliers, whereas repairable LRUs are sent to a repair shop. In the repair shop, LRUs are repaired by replacing parts that we refer to as *Shop Replaceable Units* (SRUs). SRUs, like LRUs, can be either consumable or repairable and need to be replenished from external suppliers/repair shops or an internal repair shop, respectively.

[1] Rotables are repairable parts of a system that have their own maintenance program and dedicated maintenance/overhaul capacity. Examples include aircraft engines, rolling stock bogies and weapon systems on frigates.

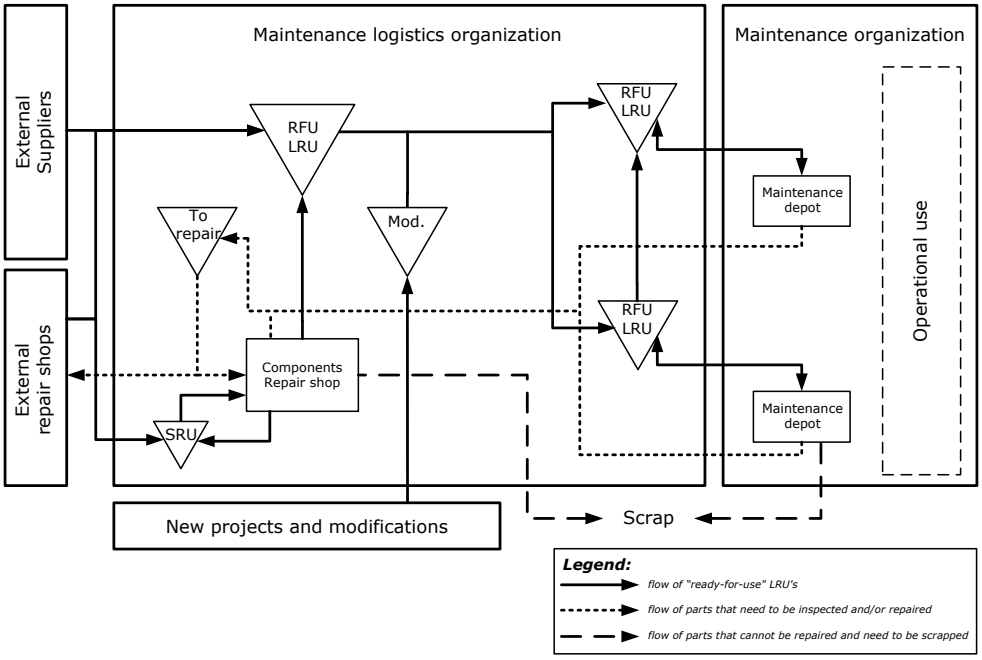

Figure 2.2 Example of a maintenance spare parts supply chain.

In general, there are multiple first level maintenance sites where assets are maintained. Associated with each site is a spare part stock point. The spare parts supply chain is a multi-echelon divergent supply chain with one or more repair shops. Furthermore, the supply chain of repairables is a closed loop system. When demand for a LRU cannot be met from local stock, emergency procedures such as lateral transshipments or emergency shipments from upstream stocking locations may be applied.

Figure 2.2 presents a typical example of a spare parts supply chain within companies that both use and maintain high value capital assets. In this and other figures throughout the thesis, upside down triangles represent stock points. A central stock point of spare parts supplies several local stock points that are incident to the first level maintenance sites. There is also a stock point of parts that still need to go to repair and a stock point of parts required for new projects and modifications that occur during the life cycle of a capital asset. In practice, these stock points are often in one and the same warehouse. For controlling the supply chain, it is convenient to consider these as separate stock points.

2.2.3 Demand characteristics of maintenance spare parts

As mentioned in §2.2.1, maintenance on a capital asset generates demand for LRUs. The MO requests the required LRUs at the MLO by creating spare parts orders. The LRUs are delivered from the stock point incident to the requesting maintenance depot. Each type of maintenance in Table 2.1 generates demand for LRUs in a different way.

Preplanned modificative maintenance generates spare parts orders some time before the planned start of the maintenance. These spare parts requirements are known and fixed ahead of time. The required LRUs are requested by the MO with a due date.

Breakdown maintenance generates demand in a unpredictable fashion. This is best modeled by stochastic models.

Demand arising from condition based maintenance is still unpredictable but only with respect to either timing (when will part demand arise?) or with respect to content (for which parts will demand arise?), but not both.

Typically, maintenance depots and MLOs make agreements on specified upper/lower bounds for key performance indicators such as

(i) The average work order delay due to unavailability of spare parts;

(ii) the percentage of work orders without delay (caused by unavailability of spare parts);

(iii) The maximum "number of unfinished work orders" due to unavailability of spare parts at any given time. Separate agreements are made on the availability of spare parts that do not cause immediate system downtime.

2.3. Framework for maintenance spare parts planning and control

In this section, we present the framework for maintenance spare parts planning and control. In Figure 2.3, an overview of processes and decisions in MLOs is presented, including their mutual connections. We separate eight different processes, which are numbered one up to eight in the figure. Within each process, we distinguish different decision levels. Decisions that are not made very frequently, i.e., once a year, are marked 'S/T' (strategic/tactical decisions); decisions made regularly, i.e. once a month or quarter, are marked 'T' (tactical decisions) and decisions made frequently, i.e., once a day/week, are marked 'O' (operational decisions).

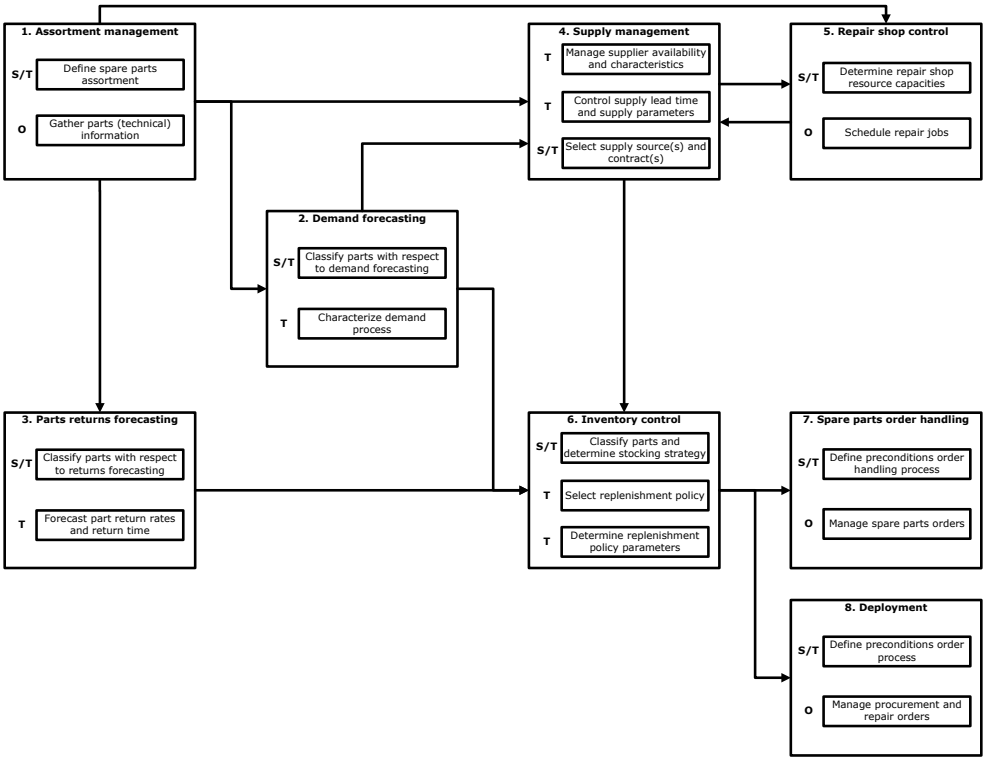

Figure 2.3 Overview of processes and decisions in maintenance logistics control.

An arc illustrates that information, e.g., data or outcomes of decisions, flows from one process to another. This information is needed to make decisions in subsequent processes. We emphasize that there are many feedback loops between the various processes. For readability, these feedback loops are left out of the figure. The framework we provide will need refinement and alterations for any particular organization and is by no means a one size fits all solution. It does however serve as a useful starting point in making specific designs of maintenance spare part planning and control systems.

2.3.1 Assortment management

Assortment management is concerned with the decision to include a spare part in the assortment and to maintain technical information of the included spare parts. We emphasize that the decision whether or not to include a part in the assortment is independent of the decision to stock the part. The process of managing the assortment

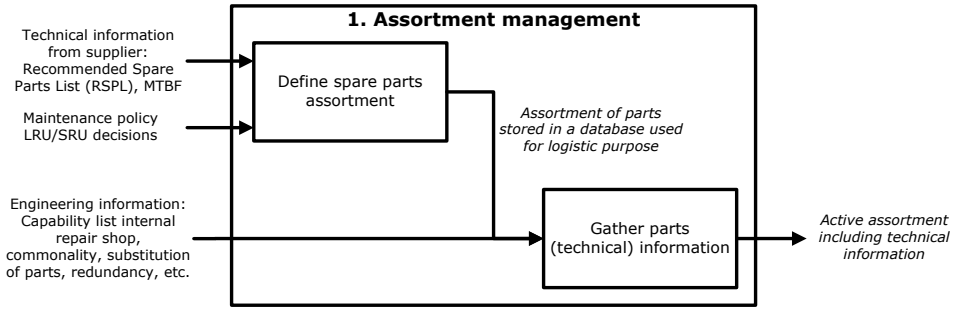

Figure 2.4 Process of managing a spare parts assortment.

can be found in Figure 2.4.

2.3.1.1 Define spare parts assortment

The decision to include (exclude) a part in (from) the assortment is usually taken shortly after procurement (phase out) of a (sub)system and strongly depends on the maintenance policy/program. Obsolete parts are excluded from the assortment. There are two options when to include a part in the assortment: before or after the first need for the part.

In case a part is included in the assortment, there is a possibility that the part is never needed during its lifecycle. In this case the time spent on collecting information, finding suppliers etc. results in unnecessary operational costs.

However, in case a part is not included in the assortment, there are two possible adverse consequences. First, when the part fails and a supplier is still available, the lead time of the part is higher due to data collection and negotiation actions. Second, when the part is needed, there may not exist any suppliers for it anymore. In this case, the part may have to be custom made. In many cases, this requires specialized technical information regarding the form, fit and function. If a part is not included in the assortment, this information is not available.

2.3.1.2 Gather parts (technical) information

Once a part is included in the assortment, (technical) information of the part is gathered and updated when necessary. The MLO needs to decide whether or not to gather and maintain parts technical information that is important for spare parts planning and control: (i) criticality, (ii) redundancy, (iii) commonality, (iv)

specificity, (v) substitution, (vi) shelf life, (vii) position in the configuration[2] and (viii) repairability. Additionally technical information regarding form, fit and function may be gathered. We also distinguish so called 'insurance' spare parts.

Parts criticality is concerned with the consequence(s) of a part failure on the asset level. Full (partial) asset breakdown means that the asset is non-operational for (a part of) all assigned use purposes. Parts that cause (partial) asset breakdown are denoted *(partially) critical*. Parts that cause no asset breakdown, i.e. the system can be used for all assigned use purposes, are denoted 'non-critical'.

Parts redundancy is the duplication of components (parts) with the intention to increase the reliability of the system. Information on parts redundancy decreases the number of stocked spare parts as it is known in advance that part failure does not cause immediate asset breakdown.

Parts commonality concerns parts that occur in the configuration of multiple assets that are maintained by the MO. For each system, the MLO needs to meet a certain service level. Information on parts commonality is needed for customer (system) service differentiation in spare parts planning as well as for the decision where to stock parts, i.e. locally or centrally.

The specificity of a part concerns the extent to which a part is tailored for and used by a customer. Parts availability at suppliers is usually low, if not zero, for specific parts and hence this might affect the size of the buffer stock needed.

Parts are substitutional in case different parts have the same form, fit and function. This means that requests for one part can be met by a substitute part. Information on parts substitution is used to prevent stocking parts for which requests can also be met by a substitute part.

The shelf life of a part is the recommended time period during which products can be stored and the quality of the parts remains acceptable for usage. This information is used to prevent stocking too many parts that are scrapped or revised after the shelf life of the part has expired.

The configuration is a list of raw materials, sub-components, components, parts and the quantities of each that are currently in an asset. Hence this list contains all the SRUs and LRUs in the system that may require maintenance during its use. The position of a part in the configuration is needed to determine at which level parts (SRUs) can be replaced, in order to repair an LRU, and what quantity of each SRU is needed. These different levels in the configuration are also called indenture levels. The initial configuration is usually provided by or available at the original equipment

[2]The configuration is similar to the Bill of Materials. However the configuration changes throughout the lifetime of an asset due to modificative maintenance whereas the bill of materials is a snapshot of the configuration at the time of initial manufacture.

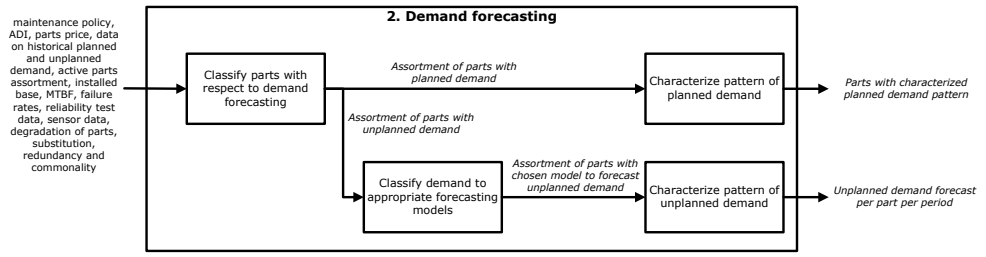

Figure 2.5 Overview of the demand forecasting process.

manufacturer (OEM) and coincides with the bill of materials.

Parts repairability concerns the identification whether a part is technically repairable and if so, whether or not the internal repair shop has the authorization (from the OEM) and the capability to repair the part. This information is needed to determine the parts' supply structure.

Technical information on form, fit and function comes in many forms depending on the technological nature of the part involved. Sometimes this information is of a sensitive nature and the OEM may charge extra for this information and/or requires non-disclosure type contracts.

'Insurance' spare parts are parts that are very reliable, highly 'critical' to asset availability and not readily available in case of failure. Often these parts are far more expensive to procure after the initial buy of the asset, compared to buying at the moment of initial asset purchase. Because of their high reliability, these spare parts often will not be used during the lifetime of the system. Example of an 'insurance' part is the propeller of a ship.

Parts (technical) information is sometimes provided by the OEM. However, it is also possible that the MLO needs to determine this technical information. All the technical information is used to improve stocking decisions and manage supply risks.

2.3.2 Demand forecasting

Demand forecasting concerns the estimation of demand for parts in the (near) future. Future demand for spare parts is either (partially) planned or unplanned and is characterized in §2.2.3. MLOs need to decide whether to use information about planned demand. The demand forecasting process is visualized in Figure 2.5.

2.3.2.1 Classify parts with respect to demand forecasting

Two types of spare parts are considered: parts for which advance demand information (ADI) is used and parts for which is it not used. Using ADI usually decreases the overall forecast error. On the other hand, it is clear that using ADI increases the difficulty, the effort and hence the operational costs to forecast demand. In case there is no information available or it is decided not to use it, then all demand is accumulated and one single demand stream is considered. Otherwise, two demand streams (planned and unplanned) are separated.

Within unplanned demand, another classification is made to aid the decision of using a particular forecasting technique. Two factors that determine what methods are appropriate are the interarrival time of demand moments and the variability of demand size. When time between demand moments is very long, then demand is said to be intermittent. When intermittence is combined with variable demand sizes, demand is said to be lumpy.

Technical information about substitution is used in forecasting to determine demand for new parts that substitute old parts. Combining demand streams, for different parts that can be met by the same spare part (i.e. substitutes), increases the overall demand forecast reliability. Technical information on commonality of new parts is used to determine how usage in different capital assets affects demand. Information on parts redundancy is used to correct the demand forecast as well.

After deciding whether or not to separate demand streams, the demand process needs to be characterized on behalf of the following three purposes: (i) to determine the number of parts to stock, (ii) to determine the repair shop capacity and (iii) to provide the necessary input for updating and characterizing supply contracts.

2.3.2.2 Characterize (partially) planned demand

Planned demand is known deterministically for the length of planning horizon. After this horizon, planned demand is unknown and has to be forecasted.

Demand for spare parts can also be partially planned in advance. Consider for example parts for which condition based maintenance using periodic inspections are applied. Parts that are needed in about $x\%$ of some types of periodic inspections are termed *x%-parts*. Combining these percentages (x) with the information on planned inspections, a forecast for this planned demand stream can be made.

Partially planned demand also occurs when condition based maintenance is applied through continuous condition monitoring. Since this is usually a rather expensive option, this is mostly only applied for a select group of expensive parts. Information

from the sensors can be used to estimate the remaining lifetime of a part and based on this the moment in time when maintenance should be conducted can be estimated with considerable more accuracy than it can be without information from sensors.

2.3.2.3 Characterize unplanned demand

Several methods are applicable to forecast unplanned demand. The first method to forecast unplanned demand is reliability based forecasting. The goal of this method is to forecast parts requirements based on part failure rates, a given installed base and operating conditions. This method determines the failure rate of one part and extrapolates the failure rate to the installed base and varying operating conditions.

The second method to forecast unplanned demand is time series based forecasting. Based on known historic requirements, extrapolations are made using statistical techniques. Examples of well-known time series based forecasting techniques are Moving averages, Smoothing methods, Croston's method and bootstrapping. The advantage of time series based forecasting is that only historical demand data is needed to forecast demand. Disadvantage is that manual changes to the demand forecasts need to be made in case the installed base or operating conditions change. The result of characterizing demand is a demand distribution per part per period.

2.3.3 Parts returns forecasting

Parts requested by the MO are sometimes returned in RFU condition. In case it is not known which part causes system breakdown, sometimes all parts that may be the cause are requested. After it is found out which part caused the breakdown, unused RFU parts are returned to the original stock point within an agreed *hand in time*. If the requested part is a repairable, a part is always returned that either (i) needs repair, (ii) is ready-for-use or (iii) is beyond repair and will be scrapped. The MLO needs to account for return rates and hand in times in their planning and control.

Consider the case of consumables. Here parts are either returned ready-for-use (with probability p_{RFU}) or not returned at all (with probability p_{con}), see also Figure 2.6. The question is now whether a part *request* should be considered a part *demand* where only part demand influences replenishment decisions. If a procurement order is placed and the part is handed in afterwards, the inventory levels grow unnecessarily.

The case of repairables is different, because replenishment orders cannot be released until a failed item is sent to the MLO. Let p_{RFU} denote the probability that a returned part is ready-for-use, p_{rep} denote the probability that a returned part needs repair and p_{con} denote the probability that a returned part will be condemned (see Figure

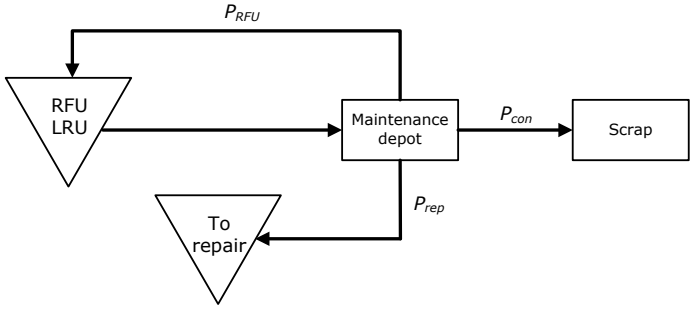

Figure 2.6 Overview of different part return streams.

2.6). These return fractions are used by inventory control in this manner: Requests for parts can be considered as demand but the lead time is altered as follows: With probability p_{RFU} the lead time is equal to the hand in time; with probability p_{rep} the lead time is the convolution of hand in time, return lead time and the repair lead time; and with probability p_{con} the lead time is the convolution of the hand in time and the procurement lead time.

The most straightforward technique to forecast return rates is to use historic return rates, possibly corrected for special events such as unusual accidents. For most parts, this technique is sufficiently accurate. For some parts, different failure modes often correspond to different types of returns. Techniques from reliability engineering can be used to estimate these return rates.

2.3.4 Supply management

Supply management concerns the process of ensuring that one or multiple supply sources are available to supply ready-for-use LRUs, as well as SRUs, at any given moment in time with predetermined supplier characteristics, such as lead time and underlying procurement contracts (price structure and order quantities). A process overview of supply management can be found in Figure 2.7.

2.3.4.1 Manage supplier availability and characteristics

The process of managing supplier availability and characteristics within MLOs is concerned with having one or more supply sources available for each spare part in the assortment, including supply characteristics. MLOs have several possible supply types: (i) internal repair shop, (ii) external repair shops, (iii) external suppliers and (iv) re-use of parts. Reuse of parts is possible by taking them out of a system at its

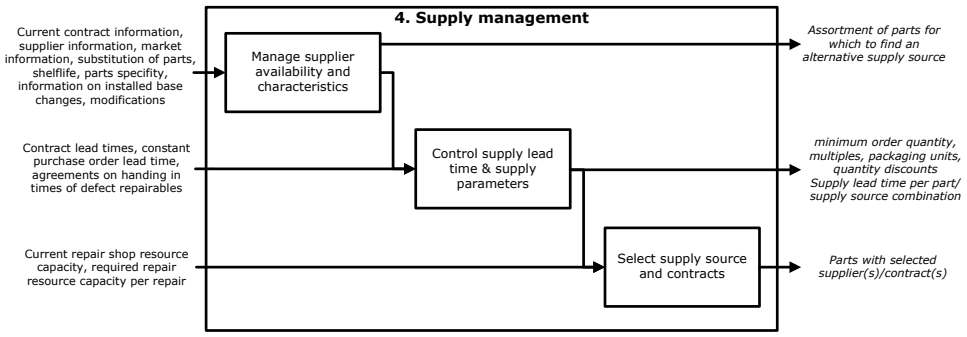

Figure 2.7 Process of managing the supply structure.

end-of-life. Within each supply type, it is possible to have multiple supply sources.

Each part has either: (i) one or more supply sources, (ii) one supply source that is known to disappear within a certain time period or (iii) no available supply source at all. In the latter case, the MLO needs to find an alternative supply source for all parts that need future resupply. Alternative supply sources are e.g. a new supplier/repair shop, a substitute part or changing the status of part from consumable to repairable (if technically possible) and contracting a repair shop to do the repair.

When the only supply source of a part is known to disappear, the MLO needs to decide whether to search for an alternative supply source or to place a final order at the current supply source. The final order decision concerns the determination of a final order quantity that should cover demand during the time no supply source will be available. The supply availability for these parts is guaranteed through the available inventory. Managing supply availability is also concerned with timely updating and maintaining current contracts with external suppliers.

MLOs also need to gather and maintain information on supply characteristics. Information concerning the following matters is needed to determine the supply lead time (distribution) and to select a (preferred) supply source and contract: (i) contractual or historical repair/new buy price(s) of the part, (ii) quantity discounts (iii) contractual lead and/or repair lead time, (iv) minimum order quantities and (v) multiples.

2.3.4.2 Control supply lead time and supply parameters

The supply lead time consists of: (i) repair or supplier lead time, (ii) procurement time, (iii) picking, transport and storage time of parts and, in case of repairables, (iv) hand in times of failed repairables. For all these components of the supply lead time,

agreements are made on planned lead times.

Using planned lead times for internal repair is justified because: (i) MLOs make agreements with the internal repair shop on planned repair lead times and (ii) the repair shop capacity is dimensioned in such a way that internal due dates are met with high reliability. Using planned lead times for external supply is justified because MLOs agree on contractual lead times with their external suppliers.

The supply lead time is determined for each part/supply source combination separately. We distinguish two types of supply lead times: (i) repair lead time and (ii) procurement lead time. For all parts that are known to be 'technically repairable', the MLO gauges the procurement lead time, the external repair lead time and the internal repair lead time, in case internal repair is possible. For consumables, only the procurement lead time needs to be gauged.

2.3.4.3 Select supply source and contracts

The MLO needs to make sure that spare parts can be replenished at any given time. For this purpose, the MLO needs to set up contracts with one or multiple supply sources in a cost efficient way. The decision is based on the following costs incurred while selecting a supply source: (i) setup and variable costs of the repair shop capability and resources, (ii) setup costs of the contract, (iii) procurement or repair costs and (iv) inventory holding costs.

The MLO uses information on supply characteristics and supply lead times to select one or more supply sources out of all possible part/supply source combinations. Important in selecting a supply source is the decision whether to designate a spare part as repairable or consumable. Alfredsson (1997) states: "The task of determining whether an item should be treated as a discardable (consumable) or repairable item is called *level-of-repair-analysis* (LORA). If the item is to be treated as a repairable item, the objective is also to determine where it should be repaired". See also Basten et al. (2009) and MIL (1993) for analogous definitions of LORA.

The MLOs should conduct a LORA that covers characteristics such as: (i) unsuccessful repairs, (ii) no-fault-found, (iii) finite resource capacities, (iv) the possibility of having multiple failure modes in one type of component, (v) the option to outsource repairs, and (vi) the possibility of pooling parts sourcing in framework agreements.

The LORA is reconsidered each year or in case of substantial changes in the asset base. The outcome of the LORA is used to reconsider the internal repair shop resource capabilities. Note that MLOs need to set up a contract for repair as well as for new buy of repairables.

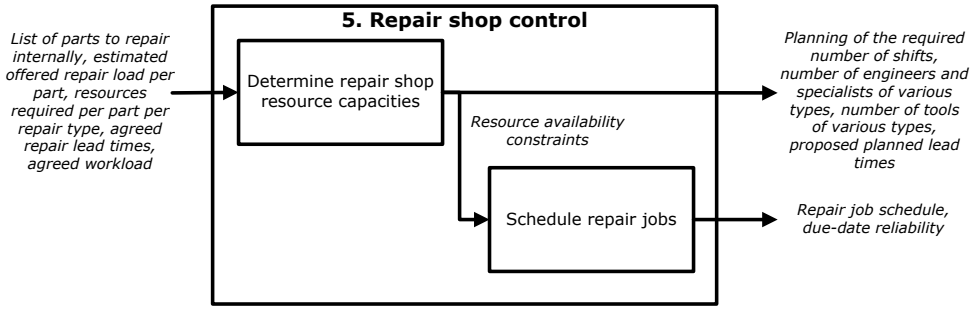

Figure 2.8 Overview of repair shop control process.

2.3.5 Repair shop control

The repair shop in the spare parts supply chain functions much like a production unit in a regular supply chain. At the interface with supply structure management, agreements are made on lead times for the repair of each LRU. Also agreements are made on the load imposed on the repair shop so that these lead times can be realized. For example, it is agreed to release no more than y parts for repair during any week.

To comply with these lead time agreements, decisions are made at a tactical and operational level. At the tactical level the capacity of the repair shop is determined and at the operational level, repair jobs are scheduled to meet their due dates. A schematic overview of repair shop control is given in Figure 2.8.

2.3.5.1 Determine repair shop resource capacities

When a repair job enters the repair shop, the sojourn time in the repair shop consists mostly of waiting time for resources such as specialists, tools and SRUs to become available. The amount of resources that are available in the repair shop determines the waiting times. These resource capacities need to be dimensioned in such a way that most repair jobs are completed within the agreed planned lead times.

Decisions need to be taken on the amount of engineers and specialists to hire, the number of shifts and the number of tools of various types to acquire. In some instances, these tools are themselves major capital investments. The SRU stocking decision lies outside the responsibility of the repair shop and is part of the total inventory control decision; see §2.3.6 for the reasoning behind this.

The resource capacity dimensioning decisions are based on the estimated repair workload, the repair workload variability (which follows from demand forecasting and parts returns forecasting) and the estimated repair time (and variability) required for

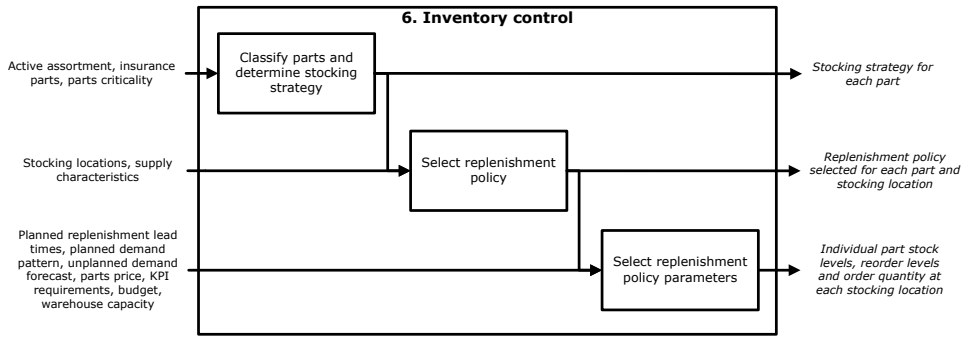

Figure 2.9 Process overview of controlling inventories.

a LRU when all resources are available. In making this decision, congestion effects need to be incorporated explicitly.

Since the costs of internal repair are mostly the result of the resources required for repair, the dimensioning decision together with the offered repair load can be combined to estimate the repair lead time and the cost of performing an internal repair. This information is used by supply structure management to periodically reconsider the LORA decision.

2.3.5.2 Schedule repair jobs

During operations, LRUs that need repair are released to the repair shop and need to be repaired within the agreed planned lead time. This naturally leads to due-dates for repair jobs. The repair job scheduling function is to schedule the repair jobs subject to the resource constraints which are a consequence of the capacity dimensioning decision. Within these constraints, specific resources are assigned to specific repair jobs for specific periods in time so as to minimize the repair job tardiness. Additionally the repair shop may batch repair jobs to use resources more efficiently by reducing set-up time and costs associated with using certain resources.

2.3.6 Inventory control

The inventory control process is concerned with the decision which spare parts to stock, at which stocking location and in what quantities. Thus, inventory planning is done centrally for all locations (multi-echelon approach). The inventory control process is visualized in Figure 2.9.

A MLO stocks LRUs in order to meet certain service levels, agreed upon with the

MO. Both the LRU and SRU inventories are centrally controlled, that is, control of SRU and LRU inventories are integrated. In this way, the multi-indenture structure of spare parts within an asset can be used in inventory control.

LRUs required for new projects and modifications are planned separately because uncertainty plays no significant role here. We will not discuss the inventory control of LRUs needed for modification in detail here.

2.3.6.1 Classify parts and determine stocking strategy

The MLO has several stocking strategies and classifies the spare parts assortment into different subsets: (i) (partially) critical spare parts and (ii) non-critical spare parts. Insurance parts are a specific subset of critical parts. The decision to stock insurance parts is not based on demand forecasts or on the contribution to a certain service level, but is based on other criteria such as supply availability, failure impact or initial versus future procurement price.

The availability of (partially) critical parts is needed to reduce system downtime. The stocking decision of (partially) critical spare parts depends on the contribution of a part to the overall service level of all (partially) critical parts. The availability of non-critical parts is needed for supporting an efficient flow of system maintenance, non-availability however does not cause immediate system downtime. Separate service level agreements are made for non-critical parts.

2.3.6.2 Select replenishment policy

The MLO is responsible for inventory replenishment of spare parts at all stocking points. To enable economies of scale in replenishment, the central warehouse replenishes the local stock points only once during a fixed period (typically a couple of days or one week). This results in a (R, S)-policy for all parts at the local stocking locations. The length of the review period is set such that internal transport of parts is set up efficiently. In order to reduce system downtime costs, it may be beneficial to use emergency shipments from the central warehouse or lateral transshipments from other local stocking locations to deliver critical parts required at a local stock point.

The MLO determines the timing and frequency of placing replenishment orders for the central stock based on supply characteristics. Spare parts for which framework-contracts are set up are usually delivered only once during a fixed period. Hence the stock level needs to be reviewed only once during this period, which results in a (R, S)-policy for these parts. The stock level of other parts is reviewed daily, resulting in an (R, s, S) or (R, s, Q)-policy for these parts.

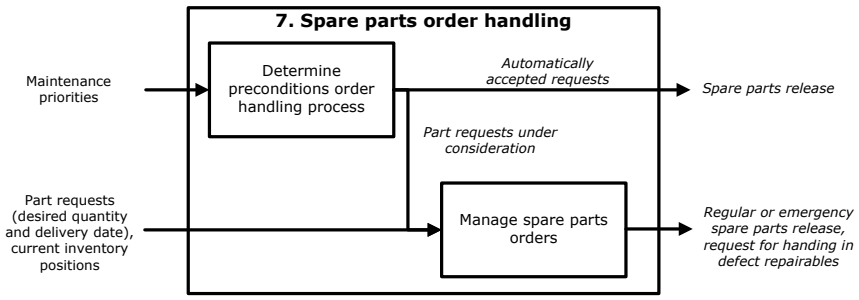

Figure 2.10 Process of handling spare parts orders.

2.3.6.3 Determine replenishment policy parameters

The MLO uses different methods to determine replenishment policy parameters for non-critical parts and (partially) critical parts. For (partially) critical parts one all encompassing model should be used to aim at a system (multi-item) service level. Optimizing policy parameters to satisfy a system service level is called the *system approach*. In maintenance logistics, this is particularly useful, because MLOs are not interested in the service level of any one part but in the amount of delay they experience in waiting for parts, regardless of which part it is specifically.

The model should contain the following characteristics: (i) multi-echelon, (ii) multi-item, (iii) multi-indenture structure, (iv) emergency shipments from central depot, (v) lateral transshipments and (vi) multiple service level criteria. Input for this model are demand forecasts and information from supply structure management (supply lead times, parts prices), parts returns forecasts and information on the current inventory positions and replenishment policies of the spare parts.

2.3.7 Spare parts order handling

As discussed in §2.2, system maintenance work order planning and release is done locally by the MOs. Each MO plans its work orders based on their available resource capacities. Resources that MOs share are spare parts. Spare parts order handling is assigned centrally to the MLO and consists of the following steps: (i) accept, adjust or reject the order, (ii) release spare parts on the order and (iii) handle return order of failed repairable(s). For each of these steps, preconditions need to be defined as well as rules to manage these steps. A process overview of handling spare parts orders is found in Figure 2.10.

2.3.7.1 Determine preconditions of the order handling process

The first decision in handling spare parts orders is to accept, adjust or reject the order. The advantage of checking spare parts orders is that unrealistic or unusual orders can be adjusted or, in case of incorrect orders, rejected. On the other hand, checking spare part orders is time consuming and increases the operational costs.

When checking spare part orders, the MLO obtains a trigger to contact the MO and adjust the order lead times and/or quantities. In this manner, MOs can reschedule certain tasks of their system maintenance work orders and adjust their spare parts orders based on the new system maintenance schedule. This might decrease system downtime (costs) caused by unavailability of spare parts.

Prioritization amongst spare parts orders while releasing spare parts is not easy in case the available stock is insufficient to meet all demand for that spare part. This is caused by the fact that the required spare parts are (i) part of a set of spare parts needed to start a maintenance task and (ii) are needed to start a different type of maintenance including different levels of criticality. Thus to fill orders, spare parts order handling faces an allocation problem similar to that found in assemble-to-order systems. The optimal solution to this problem is not generally known.

Once spare parts are released on a work order, the return process for failed repairables starts. For this purpose, the MLO creates a return order to hand in the failed repairable by the MO within the agreed hand in time.

2.3.7.2 Manage spare parts orders

Incoming spare parts orders are either automatically accepted or not, based on the preconditions set in the previous section. There might be several good reasons for unusual or unrealistic orders, hence there are no standard rules for accepting, adjusting or rejecting spare parts orders. This task lies with the MLO, who needs to consult with the MO on this.

2.3.8 Deployment

Deployment concerns the process of replenishing spare parts inventories. The deployment process consists of the following steps: (i) define preconditions of the order process and (ii) manage procurement and repair orders. A process overview of the deployment process can be found in Figure 2.11.

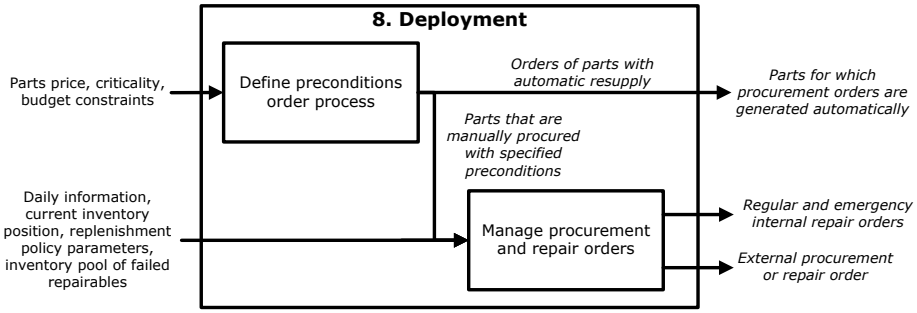

Figure 2.11 Deployment process.

2.3.8.1 Define preconditions of the order process

The replenishment policy parameters set by inventory control implicitly determine when to replenish spare parts inventories and what quantities to repair or procure. Deployment may deviate from this based on new (daily) information not known at the time the replenishment policy parameters were set, or when exceptional repair or procure orders arise from exceptional inventory levels. Deployment then starts a feedback loop to reconsider e.g. the demand forecast or supply lead times that led to this exceptional inventory level. Hence, deployment sets rules for exception management. The MLO should set a precondition on whether to replenish inventories with or without interference of deployment.

2.3.8.2 Manage procurement and repair orders

The process of managing procurement and repair orders consists of the following steps: (i) procure or request the repair of parts with the right quantity and priority, (ii) check the quality of the received spare parts and (iii) monitor supply lead times. The MLO needs to determine which quantity of each part to order and with what priority, for parts for which the procurement or repair order is checked upon release. The quantity that deployment actually orders may deviate from the order quantity set by inventory control, based on newly obtained information. When an order is received, the MLO needs to check the quality of the received parts. When orders do not arrive within the agreed lead time, deployment takes necessary recourse actions.

2.4. Framework related literature and open research topics

In this section we provide available literature that provides support for making decisions in the framework. We do this per part in the framework and in the same order as in §2.3. When discussing the literature we also identify areas that require additional research. The intention is not to provide a comprehensive or exhaustive review of the literature. Though the references we provide are a good starting point to investigate specific areas of literature in more depth and find models to support decisions that need to be made.

2.4.1 Assortment management literature

The first decision in assortment management is whether or not to include a part in the assortment. Even when no inventory will be held for a part it may be beneficial to include it in the assortment so that technical information and supply contracts are taken care of in case of a failure. For this decision we have been unable to find any literature. We propose to use simple rules of thumb based on cost and failure rates.

Most of the information gathered on parts included in the assortment has a technical character. For example the criticality of a part can be determined through a *failure mode effect and criticality analysis* (Stamatis, 1995; Ebeling, 2001). We agree with Huiskonen (2001) that these type of analyses depend on technical and not logistical part behavior.

Parts technical information can be used to decrease stock levels or manage supply risks. We have been unable to find literature that supports the decision to gather parts technical information or not. We propose to use simple rules of thumb based on cost and failure rates.

Another decision that may occur in assortment management is the decision of whether or not to include parts that can be used to serve multiple asset types, but that may be more expensive than dedicated parts. Kranenburg and Van Houtum (2007) provide a model that can serve in making this decision. Note that in making this decision the quality of parts supplied by different suppliers (in terms of reliability) should also be accounted for. If this is an important issue the model of Öner et al. (2010) can be used.

2.4.2 Demand forecasting literature

To forecast demand for spare parts traditionally two families of techniques are used, namely (i) reliability based forecasting and (ii) time series based forecasting. We would like to add a third category which we shall label (iii) maintenance planning based forecasting.

Since demand for LRUs in many cases arises due to some kind of failure of equipment, forecasting demand is equivalent to forecasting failures. A recognition of this fact leads to reliability based forecasting. The techniques from reliability engineering can be used to deal with issues such as censoring and changing operating conditions. Furthermore the forecasts obtained are related to the installed base of equipment. Thus when the installed base changes the demand forecasts can easily be updated accordingly without the need for new data. Important references for these techniques are Nelson (1982, 1990) and Ebeling (2001). More recently, reliability literature has also addressed the real time forecasting of failures using some form of degradation data from sensors. We term this *prognostics* and refer the reader to Heng et al. (2009) for a recent survey.

Time series based forecasting is the traditional technique for demand forecasting in inventory control and also finds applicability in spare part inventory control. Its use is most suited when only historic demand data is available. Many common techniques such as exponential smoothing are part of standard textbook literature on inventory control (Silver et al., 1998; Hopp and Spearman, 2001). More sophisticated techniques such as autoregressive integrated moving averages (ARIMA) can be found in the seminal work of Box and Jenkins (1970). A somewhat separate stream of literature that is especially useful in forecasting demand for spare parts was started by Croston (1972). Croston observed that demand for certain items was intermittent and spare parts typically fall into this category. To increase forecast accuracy, Croston proposes to forecast interarrival time and order quantities of demand separately. Many contributions have been made based on this idea. Teunter and Duncan (2009) benchmark many of these contributions and provide relevant references. Another technique in time series based forecasting is bootstrapping. Willemain et al. (2004) adapt this technique specifically to forecast spare part demand.

A third family of techniques that we advocate has received relatively little attention in the literature. This family of techniques bases the forecast of spare parts demand on maintenance planning information. In this manner, some demand is known exactly ahead of time. Demand for other parts may occur as a result of planned inspections. When these inspections are part of the maintenance planning, they can be used to forecast demand more accurately. As Hua et al. (2007) put it: “demand of spare part at any time is a function of equipment maintenance operations and dependent

on some explanatory variables". In particular, maintenance planning can be used to accurately forecast demand for x%-parts. This idea has found recent following in Romeijnders et al. (2012) and Wang and Syntetos (2011).

2.4.3 Parts returns forecasting literature

Parts return forecasting can be done using historic return rates. Here the methods from time series based forecasting as outlined in §2.4.2 can be used. The return rate of repairable parts that need to be scrapped can also be estimated using techniques from reliability engineering. Typically a part has several failure modes and a failure rate is associated with each failure mode. Some failure modes render the part no longer repairable while other types of failure can be repaired easily. Using models from reliability engineering (Nelson, 1982; Ebeling, 2001) these different failure rates can be estimated. Scrap rates can be determined from these estimates as the fraction of non-repairable failure rates and total failure rate. These techniques can also cope with issues such as censoring and varying operating conditions.

2.4.4 Supply management literature

When a new capital good is taken into service, supply structure planning is primarily concerned with the question which parts to designate as repairables and when an item is designated as repairable whether or not we should outsource repair. These questions are answered by a level of repair analysis (Basten et al., 2009; Barros and Riley, 2001; Alfredsson, 1997; Basten et al., 2011).

An important part in supply structure planning is setting up and maintaining relations with outside suppliers and repair shops. These issues are addressed in purchasing literature, for example the book by Van Weele (2010) covers these topics.

Another important task concerns dealing with final orders when a supplier indicates that a part will become unavailable and a final order can be placed. In case the item under consideration is a repairable handled by our own repair shop Van Kooten and Tan (2009) provide a model for decision support. Teunter and Klein Haneveld (1998) and Teunter and Fortuin (1999) provide models for decision support if it concerns consumable spare parts.

2.4.5 Repair shop control literature

In our framework we decomposed inventory control from repair shop control. Consequently the only responsibility of the repair shop is to realize certain lead times,

while inventory control is responsible to balance the workload offered to the repair shop. As such the repair shop functions much like a production unit in a conventional supply chain for which many models are available (Bertrand et al., 1990).

At the tactical level the capacity of the repair shop needs to be dimensioned. This is a machine repair problem from queueing theory (Iglehart, 1965). However it may be convenient to not directly consider the number of spare parts in the dimensioning decision in which case more general dimensioning methods may be used, e.g. from call center literature (Borst et al., 2004) or general manufacturing literature Hopp and Spearman (2001).

We note that making lead time and work load control agreements is not a simple matter. Repair capacity and inventory can both serve to buffer spare part demand variability. To find the most cost effective way to do this requires an integrated approach. To setup control and responsibilities in an organization integrating this control is not convenient. However results from models that integrate these decisions can be used to make judicious choices on lead-time agreements and work-load control. Examples include Adan et al. (2009) who show how static priorities can be used to reduce the lead times and required spare part investments for expensive parts and Hausman and Scudder (1982) who show the same result via simulation for dynamic priorities. However much useful research on this interface can still be done, and we shall return to this issue in Chapter 5 of this thesis.

For the daily scheduling of jobs many models are available (Pinedo, 2009). Also Caggiano et al. (2006) provide a model to allocate repair capacity in real time to different repair jobs based on current inventory levels. Priority schedules and use of flexible capacity in the form of overtime are discussed by Guide Jr et al. (2000), Hausman and Scudder (1982) and Tiemessen and Van Houtum (2012).

2.4.6 Inventory control literature

For parts that are not critical, usually 'regular' inventory control models can be used as they are found in standard textbooks (Silver et al., 1998; Hopp and Spearman, 2001; Zipkin, 2000). Such models include classification of parts using ABC-analysis, lot-sizing using economic order quantity (EOQ) type models and statistical inventory control.

The unavailability of a critical part leads to system downtime. Control for these parts thus becomes a paramount task. The seminal contribution in (critical) spare parts inventory control is the Multi-Echelon Technique for Recoverable Item Control (METRIC) model of Sherbrooke (1968). This model uses a multi-item approach and is valuable for controlling expensive critical parts that are replaced (mostly)

correctively. The most noteworthy contributions since METRIC are the MOD-METRIC (Muckstadt, 1973) and VARI-METRIC (Sherbrooke, 1986) extensions, that find approximate means to relax the assumptions underlying the METRIC model. MOD-METRIC and VARI-METRIC have the attractive feature of including SRUs into the analysis. The most important models in spare parts inventory control have been consolidated in the books by Sherbrooke (2004) and Muckstadt (2005). Guide Jr. and Srivastava (1997) and Kennedy et al. (2002) provide literature overviews on spare part inventories and issues surrounding them such as emergency procedures (Alfredsson and Verrijdt, 1999; Song and Zipkin, 2009), lateral transshipments (Paterson et al., 2011), interaction with finite repair capacity (Sleptchenko et al., 2005), interaction with maintenance policies and obsolescence.

A new aspect in spare parts inventory control that literature has not yet addressed is advance demand information through prognostics, maintenance planning or a combination of these two. Most of the literature on spare parts inventory control assumes demand for parts arises from corrective maintenance, i.e. failures of parts. In the environment we consider, most maintenance is either condition or usage based. Thus more sophisticated demand models that leverage the availability of information regarding maintenance are of interest.

While inventory models with more sophisticated demand models have been studied for regular supply chains, this knowledge is not immediately transferable to spare part supply chains. The main reason for this is that repairables have a closed loop supply chain such that the repair of a part cannot start before its replacement.

2.4.7 Spare parts order handling literature

When a spare part order has been accepted, the part may not be on stock locally. The priority that one may give to alternate sources such as other local stock-points or the central warehouse is an important issue. Current literature (Alfredsson and Verrijdt, 1999) usually assumes local transshipments are favored over emergency shipments from the central warehouse. In the present context this assumption is often violated, probably with good reason. Literature has yet to investigate this.

The second issue concerns allocation of spare parts to work orders. We already pointed out that this allocation problem is similar to the one found in *assemble-to-order* (ATO) systems. This problem is known in general to be NP-hard (Akçay and Xu, 2004) even without the stochasticity involved in demand. Also in the present context the question of when, whether and how to hold back spare parts in order to fill complete work orders is still an open question for which only limited results are available (Lu et al., 2010). These limited results suggest that it can be beneficial to hold back inventory to fill complete work-orders. We note also that it has been pointed out that this allocation

decision should be jointly optimized with the inventory control decision (Akçay and Xu, 2004).

2.4.8 Literature overview

Table 2.2 gives a brief overview of the literature we have discussed in this section. The first column of Table 2.2 contains the processes in the framework. The second and third column contain the specific topic related to the decision function and the relevant references respectively.

Two review papers have been written on spare part management (Kennedy et al., 2002; Guide Jr. and Srivastava, 1997). These papers provide an enumerative review of the state-of-the-art at the time of writing. In this section, we do not attempt to provide an exhaustive review of contributions on these subjects. Rather, we provide relevant references for each part of the decision framework to facilitate decision making.

Within the literature some interesting topics remain. Here we describe two topics that we shall return to in Chapters 4 and 5.

Most forecasting methods, also for spare parts, are solely based on analyzing past observations. Demand for maintenance spare parts arises out of maintenance. Recent developments in *condition based maintenance* (CBM) and remote monitoring (e.g. Wang and Syntetos, 2011; Heng et al., 2009) enable us to predict the need for maintenance more carefully and in realtime. The ramifications for spare parts demand modeling are not fully understood and deserve further investigation.

On the inventory control side, there are also challenges associated with this. Most spare parts inventory models assume Poisson demand. When forecasts evolve in realtime based on sensor information, this assumption is not tenable. In essence, the information from prognostics offer some kind of advance demand information. Leveraging this information in inventory control is not straightforward. Consider repairables; it is usually not possible to react to this realtime information by changing the number of spare repairables. The repair lead time of different repairable items perhaps can be influenced more or less in realtime, thus leveraging advance demand information from prognostics. How to organize this efficiently is an open research topic.

Table 2.2 Literature on different decision functions in the framework.

Process	Topic(s)	Literature
1. Assortment Management	FME(C)A (failure mode effects (criticality) analysis)	Stamatis (1995) Ebeling (2001)
	Criticality, specificity, value	Huiskonen (2001)
	Commonality	Kranenburg and Van Houtum (2007)
	Reliability and quality	Öner et al. (2010)
2. Demand forecasting	Overview	Altay and Litteral (2011)
	Time series analysis	Box and Jenkins (1970) Chatfield (2004)
	Bootstrapping	Willemain et al. (2004)
	Croston methods	Croston (1972) Teunter and Duncan (2009)
	Life data analysis of equipment	Nelson (1982); Ebeling (2001)
	Prognostics	Heng et al. (2009)
	Linking forecasting to maintenance planning	Wang and Syntetos (2011) Hua et al. (2007) Romeijnders et al. (2012)
3. Parts return forecasting	Scrap rates/Reliability engineering	Nelson (1982); Ebeling (2001)
4. Supply management	LORA (level of repair analysis)	Basten et al. (2009, 2011) Alfredsson (1997) Barros and Riley (2001)
	Contract management	Van Weele (2010)
	Last buy	Van Kooten and Tan (2009) Bradley and Guerrero (2009) Teunter and Klein Haneveld (1998) Teunter and Fortuin (1999)
5. Repair shop control	Capacity dimensioning and machine repairman models	Iglehart (1965); Borst et al. (2004) Chakravarthy and Agarwal (2003)
	Scheduling / Capacity assignment	Caggiano et al. (2006); Pinedo (2009)
	Overtime usage	Scudder (1985) Scudder and Chua (1987)
	Priority assignment to jobs	Scudder (1986); Guide Jr et al. (2000) Adan et al. (2009) Tiemessen and Van Houtum (2012) Hausman and Scudder (1982)
6. Inventory control	Review articles and books	Sherbrooke (2004); Muckstadt (2005) Kennedy et al. (2002) Guide Jr. and Srivastava (1997)
	METRIC-type models (multi-echelon technique for recoverable item control)	Sherbrooke (1968, 1986) Muckstadt (1973); Graves (1985)
	Lateral transshipments	Lee (1987); Paterson et al. (2011) Kranenburg and Van Houtum (2009)
	Emergency procedures	Alfredsson and Verrijdt (1999) Verrijdt et al. (1998)
	Finite repair capacity	Sleptchenko et al. (2002) Díaz and Fu (1997); Adan et al. (2009) Caggiano et al. (2006)
7. Spare parts order handling	Allocation policies	Akçay and Xu (2004); Lu et al. (2010)
8. Deployment literature	Behavioral aspects of planning	Fransoo and Wiers (2006); Wiers (2009) Fransoo and Wiers (2008)

2.5. Concluding remarks

In this chapter, we presented a framework for maintenance spare parts planning and control for organizations that use and maintain high-value capital assets. This framework can be used to increase the efficiency, consistency and sustainability of decisions on how to plan and control a spare parts supply chain. The applicability and benefits of our framework are demonstrated through a case study at NedTrain, a company that maintains rolling stock. We also provided literature to assist in decision making for different parts of the framework and identified open research topics.

Chapter 3

Rotable overhaul and supply chain planning

"In preparing for battle I have always found that plans are useless, but planning is indispensable."

Dwight E. Eisenhower

3.1. Introduction

The availability of capital assets is crucial to keep the primary processes of their owners up and running. While the acquisition cost of capital assets is substantial, the costs associated with maintenance and downtime over the lifetime of the asset is typically 3 to 4 times the acquisition price, even when the future costs of maintenance and downtime are discounted (Öner et al., 2007). Accordingly, there has been much focus and research on what is called life cycle costing (LCC), see Gupta and Chow (1985) and Asiedu and Gu (1998). The LCC approach to decision making in asset acquisition, maintenance, and disposal stipulates that the consequences of decisions should be accounted for over the entire lifetime of the asset in question.

Another factor influencing maintenance is the modular design of many technical systems. Usually, a capital asset is not maintained in its entirety at any one time. Instead, different modules of the system are dismounted from the asset and replaced by *ready-for-use* modules. After replacement, the module can be overhauled while the

capital asset is up and running again. Exchanging modules, rather than maintaining them on the spot, increases the availability of capital assets, as assets are only down for the time it takes to replace a module. After overhaul, the module is ready-for-use again and can be used in a similar replacement procedure for another asset. To make this system work, some spare modules are needed, and they form a so called *turn-around stock*.

In this chapter, we consider the replacement of modules that have their own maintenance program. The maintenance program stipulates a maximum amount of time/usage a module is allowed to be operational before it needs to be overhauled. We refer to this time allowance as the *maximum inter overhaul time* (MIOT), and we assume that there is a direct relation between the time a module has been in the field and its usage. Due to safety regulations, or contracts with the original equipment manufacturer (OEM), the MIOT is usually quite conservative and so most modules are almost exclusively maintained preventively. We call the practice described in the previous paragraph as *maintenance-by-replacement*. Note that this is similar to, but different from repair-by-replacement wherein components are replaced for unplanned corrective or condition based maintenance, as opposed to planned usage-based preventive maintenance. We refer to the modules involved as *rotables*, because they rotate through a closed-loop supply chain. At this point, we emphasize that rotables differ from repairables as they are studied in much of the spare parts inventory control literature (e.g., Sherbrooke, 2004). Repairables do not have a maintenance program of their own, and consequently, the need for replacement of repairables is usually characterized by stochastic models such as the (compound or Markov modulated) Poisson process. By contrast, rotables do have their own maintenance program, and so replacements and overhauls of rotables are planned explicitly by a decision maker.

This chapter is motivated by a maintenance-by-replacement system in place at NedTrain, a Dutch company that performs maintenance of rolling stock for several operators on the Dutch railway network. However, the model is generic for companies with a maintenance-by-replacement system such as airlines that maintain engines by replacement. Below, we describe several characteristics and constraints of maintenance-by-replacement systems and their implications for planning.

In a maintenance-by-replacement system, replacements and overhauls are subject to the following two constraints. A replacement may not occur, unless a ready-for-use rotable is available to replace the rotable that requires overhaul, so that the asset can immediately return to operational condition. An overhaul cannot occur, unless there is available capacity in the overhaul workshop. Since the result of an overhaul is a ready-for-use rotable, these constraints are connected.

The maintenance programs of rotables also impose constraints on a maintenance-by-

replacement system. For each rotable type, the maintenance program stipulates a MIOT, the maximum amount of time a rotable is allowed to be operational before it needs to be overhauled. Note that the decision to replace a rotable in some period t directly implies that the replacing rotable needs to be replaced before time $t + \text{MIOT}$.

With respect to the timing of rotable overhauls and replacements, the LCC perspective offers opportunities. In traditional maintenance models, the focus is on postponing maintenance as long as possible, thereby taking advantage of the technical life of the unit to be maintained. This approach does not necessarily lead to optimal decisions over finite lifetimes of assets. To see why, consider the following example based on practice at NedTrain. The typical lifetime of a rolling stock unit is 30 years. Bogies are important rotables in a train, with MIOTs that range from 4 to 10 years. Suppose the MIOT of two types of bogies is 7 years, and both types of bogies belong to the same type of train. Then, if replacements are planned to occur just in time, bogie replacements occur 4 times during the life cycle of this train type, namely in years 7, 14, 21, and 28. Another plan, that is feasible with respect to overhaul-deadlines, is to replace in years 6, 12, 19, and 25. Note that it is possible to replace rotables earlier than technically necessary, i.e. throwing away some of the useful life of the equipment, *without* increasing the number of replacements (and overhauls) that are needed during the lifetime of an asset. To smoothen the workload of the overhaul workshop, it may be possible to overhaul the first type of rotable according to the first schedule, and the second type of rotable according the second. In general, the flexibility in the exact timing of replacements and overhauls can be used to smoothen the workload of the overhaul workshop and utilize other resources more efficiently *without* losing efficiency by throwing away remaining useful life of rotables. In effect, we are not and should not be concerned with minimizing the amount of useful lifetime on rotables that is wasted in the short run. Rather, we should minimize the cost of maintenance and overhaul that rotables incur over the lifetime of the asset they serve, which is finite. The renewal reward theorem (e.g. Ross, 1996) that has proven beneficial in many reliability and maintenance engineering applications (e.g. Ebeling, 2001) cannot be applied in this setting. The reason for this is that the horizon we consider is not infinite (not even by approximation). To see this, consider again the example above. Only a few renewals (4 in the example) occur during the time a rotable is in the field, and the last renewal has very different characteristics from the other renewals in that the last renewal ends with replacing the asset for which the rotable is used, rather than overhauling the rotable itself.

In this chapter, we study a periodic planning model for the aggregate planning of rotable replacements and overhaul for multiple rotable types that use the same resources in an overhaul workshop. In each period, decisions need to be made regarding:

- How many rotables of each type to replace;
- How many replaced rotables of each type to release to the revision work shop;
- How many rotables to buy for new assets entering the field.
- How to change the capacity levels in the revision work shop.

Replacement decisions are subject to the availability of ready-for-use rotables to complete the replacement. The release of replaced rotables to the revision work shop is subject to capacity constraints. Finally, changing capacity levels in the work shop is constrained relative to current capacity levels. We take the LCC perspective by taking the finite life cycle of assets into consideration. The model we present should be implemented in a rolling horizon, i.e., the model generates decisions for the next 30 or so years, but only the decisions for the coming few months say should be implemented. As time progresses, estimates of input parameters for our model become more accurate, and the model should be solved again to generate decisions that are based on these more accurate estimates.

This chapter is structured as follows. In §3.2, we review the literature on maintenance and aggregate supply chain planning. We provide and analyze our model in §3.3. Computational results based on a real life case are presented in §3.4. In §3.5, we present computational results for a large test bed of randomly generated instances. Finally, conclusions are offered in §3.6.

3.2. Literature review and contribution

Aggregate planning is performed in many contexts and businesses. We review the literature on maintenance planning in §3.2.1. Since our model also deals with the rotable supply chain, we review aggregate planning models in the context of production and supply chain in §3.2.2. In §3.2.3, we explain our contribution relative to the literature discussed.

3.2.1 Preventive maintenance and capacity planning

Wagner et al. (1964) are among the first to consider the joint problem of preventive maintenance and capacity planning. They consider a setting where a set of preventive maintenance tasks is to be planned, while fluctuations in work-force utilization are to be kept at a minimum. The objective is approximately met by formulating the problem as a binary integer program and using rounding procedures to find feasible solutions. Paz and Leigh (1994) give an overview of many different issues involved

with maintenance planning and review much of the literature from before 1993. They identify manpower as the critical resource that has to be reckoned with in maintenance planning.

More recent research on maintenance planning includes Charest and Ferland (1993), Chen et al. (2010), Safaei et al. (2011), and Cho (2011). Cho (2011) formulate a mixed-integer program (MIP) to schedule both the usage and maintenance of individual aircraft in a military application. The objective in their model is to smooth the workload for maintenance personnel as much as possible. Safaei et al. (2011) consider short term maintenance scheduling to maximize the availability of military aircraft for the required flying program. The problem is cast as a (MIP) in which the required workforce is the most important constraint. Chen et al. (2010) study short-term manpower planning using stochastic programming techniques and apply their model to carriage maintenance at the mass rapid transit system of Taipei. The horizon they consider is around a week and their model allows for random maintenance requirements due to break-down-maintenance (as opposed to planned preventive maintenance). Charest and Ferland (1993) study preventive maintenance scheduling where each unit that is to be maintained is fixed to a rigid maintenance schedule with fixed inter-maintenance intervals. They model the problem as a MIP and solve this MIP with various heuristic methods such as exchange procedures and tabu search.

A closely related problem is the clustering of frequency constrained maintenance activities when a set-up cost is associated with performing maintenance. Van Dijkhuizen and Van Harten (1997) study a model where a fixed set-up cost can be shared by clustering maintenance activities. Under this assumption they provide a polynomial time dynamic programming algorithm. Zarybnisky (2011) consider a richer cost structure in which the set-up cost of clustering maintenance depends on the disassembly sequence needed to perform all the maintenance activities. They provide two approximation algorithms that compute cyclic maintenance schedules with guaranteed performance factors of 2 and $1/\ln(2) = 1.4427$ respectively.

Recently, some attention has also been paid to the availability of ready-for-use rotables as a critical constraint in maintenance planning. Joo (2009) explicitly considers the availability of ready-for-use rotables as an essential constraint in their overhaul planning model. Joo (2009) considers a set of rotables of a single type that has to meet an overhaul deadline in the (near) future. The model is set up such that overhaul is performed as late as possible, but before the deadline and within capacity constraints. The key idea is that the useful life of a rotable must be used to the fullest extent possible. Joo (2009) uses a recursive scheme to plan rotable overhaul that is very much akin to dynamic programming.

3.2.2 Aggregate production and supply chain planning

Aggregate planning in production environments was first proposed by Bitran and Hax (1977), and has been expanded upon by many authors (e.g. Bitran et al. (1981, 1982)). Today, aggregate production planning models have found their way into standard textbooks in operations and production management (e.g. Silver et al. (1998), Hopp and Spearman (2001), Nahmias (2009)). These aggregate production planning (APP) models are used to plan workforce capacity and production quantities of product families over several periods. Similar models are also used in supply chain planning. These models are described and reviewed in Billington et al. (1983), Erengüç et al. (1999), De Kok and Fransoo (2003) and Spitter et al. (2005). Although all these models generate a production plan for several periods into the future, it is understood, that only the decisions for the upcoming period should be implemented. After this period lapses, new information becomes available and existing information becomes known more accurately. In this new situation the model is rerun to generate decisions for the next period. The reason to include many periods in the model is to be able to evaluate the impact of the decision in the current period further into the future. This way of working is called rolling horizon planning.

Aggregate maintenance planning differs from aggregate production planning in two fundamental ways. First, while in APP exogenous demand triggers the use of production capacity either implicitly or explicitly, maintenance requirements are necessarily endogenous to the modeling approach. The reason for this is that preventive maintenance needs to be performed within limited time intervals due to safety and/or other reasons. Thus, a decision to maintain a rotable at some time t also dictates that the replacing rotable itself be replaced before time t+MIOT. Here too, the LCC perspective has an added value. While MIOTs have to be respected, there is considerable freedom in the exact timing of performing maintenance without increasing the number of times that preventive maintenance is performed during the life cycle of an asset. This flexibility however, can only be leveraged by considering the entire life cycle in the planning process. When this is done, flexibility can be used to utilize resources such as workforce and turn-around stock efficiently. We already noted that rolling horizon planning considers the impact of decisions in the current period to costs in future periods. In the case of maintenance planning, the relevant planning horizon is the lifetime of the assets that are to be maintained.

Second, maintenance has a fundamentally different capacity restriction in the availability of rotables for replacement actions. While production capacity levels are not directly influenced by earlier production quantities, the availability of ready-for-use rotables depends on the number of rotables that have undergone overhaul in previous periods. Thus the number of rotables in the closed-loop supply chain form a special type of capacity constraint. For a recent literature review on closed-loop supply

chains, see Ilgin and Gupta (2010). A fundamental difference between the closed-loop supply chain studied in this paper and other closed-loop supply chains studied in literature so far, is that in this case a return (replacement) automatically generates another return within some preset fixed maximum period of time, the MIOT.

3.2.3 Contribution

In the field of planned preventive maintenance, our model has several contributions to existing literature that we summarize below:

(a) Our model can be used for tactical decision making in which the effects of decisions over long horizons need to be considered. These long horizons explicitly incorporate LCC considerations into decision making and utilize the flexibility there is with respect to the exact timing of overhauls over the whole life cycle of an asset. However, we do not propose to fix a plan for very long horizons; we do propose accounting for consequences of decisions over long horizons.

(b) Our model makes the constraints imposed by a finite rotable turn-around stock explicit by modeling the rotable supply chain. It also supports the decisions regarding the size of rotable turn-around-stocks.

(c) Our model considers multiple rotable types that utilize the same overhaul capacity. For each rotable type, the model plans multiple overhauls into the future.

(d) We perform a case study, and show that a linear programming relaxation of our optimization problem yields sufficiently accurate results to aid in decision making. We also provide useful insights about planning for NedTrain, the company involved in the case study. In a numerical experiment where instances are generated randomly, we show that the solution to the LP relaxation is usually sufficiently accurate to aid decision making.

3.3. Model

We consider an installed base of capital assets and a supply chain of rotables in a maintenance-by-replacement system. The rotables in this supply chain go through the same overhaul workshop and their overhaul requires the availability of a fixed amount of resources in the overhaul workshop. Each asset consists of several rotables of possibly different type. For each rotable type, there is a population of this rotable type in the field. Each rotable in the population of a type requires overhaul before

Table 3.1 Example of regular and aggregated time periods and the set T_y^Y

Time in aggregated periods (Y)	1				2				3			
	T_1^Y				T_2^Y				T_3^Y			
Time in periods (T)	1	2	3	4	5	6	7	8	9	10	11	12

its MIOT has lapsed since the rotable has gone into active use. For the aggregate planning problem under consideration, we divide time in periods. We let T denote the set of periods in the planning horizon, $T = \{1, \ldots, |T|\}$. The length of a period is typically one month while the length of the planning horizon should be at least the length of the life cycle of the assets in which the rotables function. In this way, the model can capture the entire LCC. For rolling stock and aircraft, this planning horizon is about 25-35 years. We let I denote the set of different types of rotables. The first (last) period in the planning horizon during which rotables of type $i \in I$, are in the field is denoted a_i (p_i), $a_i < p_i$. For most types of rotables $a_i = 1$, meaning that rotables of type i are already in the field when a plan is generated. Rotables always support assets and companies plan the disposal of these assets, as well as their replacement with a newer version. When $a_i > 1$, type i rotables support an asset which the company plans to start using in period a_i. Similarly, when $p_i < |T|$, rotable type i supports an asset that will be disposed of in period p_i. We let $T_i^I = \{a_i, \ldots, p_i\}$ denote the set of periods in the planning horizon during which rotables of type $i \in I$ are active in the field. Furthermore, we let I_t denote the set of rotables that are active in the field during period $t \in T$: $I_t = \{i \in I | a_i \leq t \leq p_i\}$.

We also define a set of aggregated periods, $Y = \{1, \ldots, |Y|\}$. Typically an aggregated period is a year. Furthermore, we let T_y^Y denote the set of periods that are contained in the aggregated period $y \in Y$. Table 3.1 shows an example of how T, Y and T_y^Y relate to each other. The example concerns a horizon of three aggregated periods (e.g. years) and 12 regular periods (e.g. quarters). T_1^Y contains the periods contained in the first aggregated period (e.g. the quarters of the first year).

In the rest of this section we will describe the equations that govern different parts of the system under study.

3.3.1 Supply chain dynamics

The rotable supply chain is a two-level closed-loop supply chain as depicted in Figure 3.1. There are two stock-points where inventory of rotables that are ready-for-use and rotables requiring overhaul, respectively, are kept.

We let the variables $B_{i,t}$ ($H_{i,t}$) denote the number of ready-for-use (overhaul requiring)

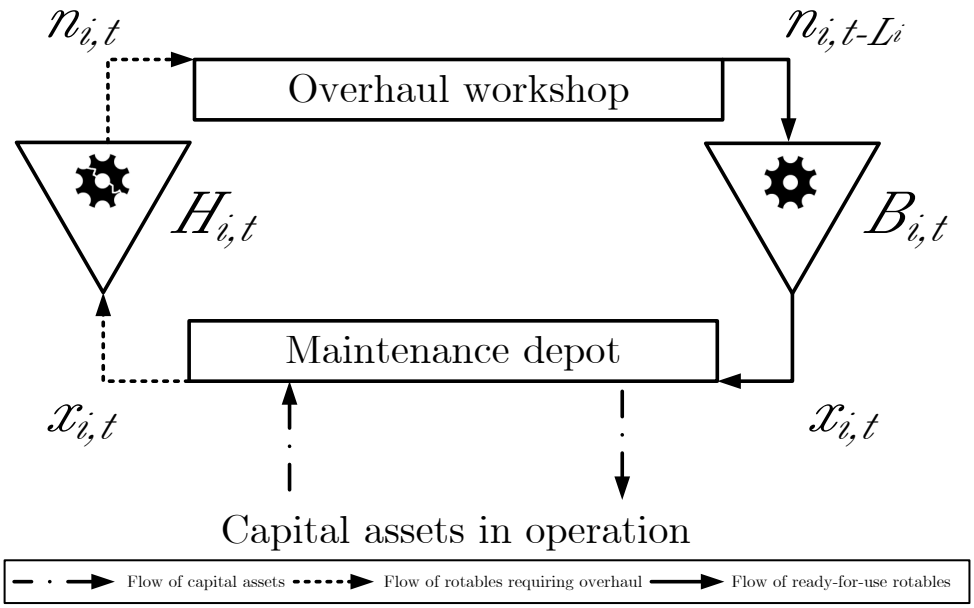

Figure 3.1 Rotable supply chain overview.

rotables of type $i \in I$ in inventory *at the beginning* of period $t \in T_i^I$. We let the decision variable $x_{i,t}$ denote replacements of rotables of type $i \in I$ *during* period $t \in T_i^I$. We assume that the time required to replace a rotable is negligible compared to the length of a period. The overhaul workshop acts as a production unit as defined in supply chain literature (De Kok and Fransoo, 2003). This means that when an overhaul order is released at any time t, the rotable becomes available ready-for-use at time $t + L_i$. Thus, L_i is the overhaul lead time and we assume it is an integer multiple of the period length considered in the problem. We let the decision variable $n_{i,t}$ denote the number of overhaul orders for rotables of type $i \in I$ released in the course of period t. The supply chain dynamics are described by the inventory balance equations:

$$B_{i,t} = B_{i,t-1} - x_{i,t-1} + n_{i,t-L_i-1}, \qquad \forall i \in I, \quad \forall t \in T_i^I \setminus \{a_i\} \tag{3.1}$$

$$H_{i,t} = H_{i,t-1} + x_{i,t-1} - n_{i,t-1}, \qquad \forall i \in I, \quad \forall t \in T_i^I \setminus \{a_i\}. \tag{3.2}$$

Equations (3.1) and (3.2) require initial conditions. The stock levels for rotables already in the field in the first planning period ($a_i = 1$) are initialized by the parameters B_i^{d} and H_i^{d} respectively; so $B_{i,a_i} = B_i^{\mathrm{d}}$ and $H_{i,a_i} = H_i^{\mathrm{d}}$ if $a_i = 1$. Here, and throughout the remainder of this chapter, the superscript d is used for parameters known from data that initialize variables. (Note that B_{i,a_i} is a variable and B_i^{d} is a parameter known from data.) For rotables that enter the field after

the first period, the initial stock level conditions are to start with the entire turn-around stock $S_i \in \mathbb{N}$ consisting of ready-for-use repairables, and no rotables requiring maintenance; so $B_{i,a_i} = S_i$ and $H_{i,t} = 0$ if $a_i > 1$. The initial turn-around stock levels for rotables that are not yet in the field in period 1, S_i, are decision variables. For $t = a_i - L_i + 1, \ldots, a_i - 1$, $n_{i,t}$ also has initial conditions: $n_{i,t} = n^{\mathrm{d}}_{i,t}$ for $t \in \{a_i - L_i + 1, \ldots, a_i - 1\}$. These initial conditions are known from data if $a_i = 1$ and set to 0 if $a_i > 1$. We assume that when $n_{i,t}$ overhaul orders are released during period t, these releases occur uniformly during that period.

3.3.2 Workforce capacity and flexibility in the overhaul workshop

The workforce capacity in the workshop is flexible. Workforce is acquired or disposed of at the ending of each aggregated time period $y \in Y$. We let the decision variable W_y denote the available labor hours during aggregated period $y \in Y$. For example, if the length of an aggregated period is a year, W_y represents the number of labor hours to be worked during that year given the number of contracts with laborers. However, there is flexibility as to when exactly these hours are to be used during the aggregated period (year). If we let the decision variable w_t denote the amount of labor hours used during period $t \in T$, this flexibility can be expressed as follows:

$$W_y = \sum_{t \in T^Y_y} w_t, \quad \forall y \in Y. \tag{3.3}$$

The average number of hours worked during any period $t \in T^Y_y$ is $W_y/|T^Y_y|$. We let the parameters δ^l_t and δ^u_t denote lower and upper bounds on the fraction of $W_y/|T^Y_y|$ that is utilized during period $t \in T^Y_y$:

$$\delta^l_t W_y/|T^Y_y| \le w_t \le \delta^u_t W_y/|T^Y_y|, \quad \forall y \in Y, \quad \forall t \in T^Y_y. \tag{3.4}$$

Thus the flexibility of manpower planning per period is also constrained by (3.4).

The labor allocated in any period t affects possible overhaul order releases as follows. We let r_i denote the amount of labor hours required to start overhaul of a type $i \in I$ rotable. Then overhaul order releases must satisfy:

$$\textstyle\sum_{i \in I_t} r_i n_{i,t} \le w_t, \quad \forall t \in T. \tag{3.5}$$

Finally, we note that W_y can be changed from one aggregated period to the next. Such a change from aggregated period y to $y + 1$ is bounded from below and above as a fraction of W_y by Δ^l_y and Δ^u_y respectively:

$$\Delta^l_y W_y \le W_{y+1} \le \Delta^u_y W_y, \quad \forall y \in \{1, \ldots, |Y| - 1\}. \tag{3.6}$$

Finally, we note that W_1 is initialized by the parameter W^{d}.

3.3.3 Rotable availability

Since the asset from which the rotables are to be replaced requires high availability, replacements may not occur unless there is a ready-for-use rotable available to complete the replacement. Similarly, we require that the release of an overhaul order must be accompanied immediately by a rotable in need of overhaul. Recalling our assumption that the replacements and overhaul order releases during any period are uniformly distributed over that period, rotable availability can be expressed as

$$n_{i,t} \leq H_{i,t} + x_{i,t}, \qquad \forall i \in I, \quad \forall t \in T_i^I, \tag{3.7}$$

$$x_{i,t} \leq B_{i,t} + n_{i,t-L_i} \qquad \forall i \in I, \quad \forall t \in T_i^I. \tag{3.8}$$

3.3.4 Overhaul deadlines propagation

Due to safety and possibly other reasons, the maintenance program of rotables of type $i \in I$ stipulates that any rotable of type i has to be replaced before it has been operational for q_i periods. Thus for each period in the planning horizon, there are a number of rotables of type $i \in I$ that have to be replaced before or in that period and we denote this quantity $D_{i,t}$ for rotables of type $i \in I$ in period $t \in T_i^I$. For a given rotable type $i \in I$, these quantities are known for period a_i up to $\min\{a_i + q_i - 1, p_i\}$ and given by $D_{i,t}^{\mathrm{d}}$.

We assume that rotables of any type are replaced in an oldest rotable first fashion, i.e., whenever a rotable of any type is to be overhauled, the specific rotable of that type that has been in the field the longest is always chosen. Thus, from period $a_i + q_i$ onwards

$$D_{i,t} = x_{i,t-q_i}, \quad \forall i \in I, \quad \forall t \in \{a_i + q_i, \ldots, p_i\}. \tag{3.9}$$

It is possible to replace rotables ahead of time, and we let U_i^{d} denote the number of rotables of type $i \in I$ that have been replaced ahead of time at time $a_i - 1$. To comply with the maintenance program, the replacements have to satisfy:

$$U_i^{\mathrm{d}} + \sum_{t'=a_i}^{t} x_{i,t'} \geq \sum_{t'=a_i}^{t} D_{i,t'} \quad \forall i \in I, \forall t \in T_i^I. \tag{3.10}$$

This constraint can also be described using an auxiliary variable, $U_{i,t}$, that represents the number of replacements of rotables of type i in excess of what is strictly necessary by period t.

Proposition 3.1 *The set of inequalities* (3.10) *is equivalent to the set of constraints*

$$x_{i,t} \geq D_{i,t} - U_{i,t-1}, \qquad \forall i \in I, \quad \forall t \in T_i^I, \tag{3.11}$$

$$U_{i,t} = x_{i,t} - D_{i,t} + U_{i,t-1}, \qquad \forall i \in I, \quad \forall t \in T_i^I \setminus \{p_i\} \tag{3.12}$$

$$U_{i,a_i-1} = U_i^d, \qquad \forall i \in I. \tag{3.13}$$

PROOF: We show that (3.12)-(3.13) imply that

$$U_{i,t} = U_i^{\mathrm{d}} + \sum_{t'=a_i}^{t} x_{i,t'} - \sum_{t'=a_i}^{t} D_{i,t'}, \quad \forall i \in I, \quad \forall t \in \{a_i - 1, \ldots, p_i - 1\}. \tag{3.14}$$

Substituting (3.14) back into (3.11) yields (3.10). To verify that (3.14) and (3.12)-(3.13) are equivalent, we use induction. First observe that (3.13) implies that (3.14) holds for all $i \in I$ and $t = a_i - 1$. Now suppose that (3.14) holds for some $i \in I$ and $t - 1 \in \{a_i - 1, \ldots, p_i - 1\}$. Then (3.12) implies that

$$\begin{aligned} U_{i,t} &= x_{i,t} - D_{i,t} + U_{i,t-1} \\ &= x_{i,t} - D_{i,t} + U_i^{\mathrm{d}} + \sum_{t'=a_i}^{t-1} x_{i,t'} - \sum_{t'=a_i}^{t-1} D_{i,t'} \\ &= U_i^{\mathrm{d}} + \sum_{t'=a_i}^{t} x_{i,t'} - \sum_{t'=a_i}^{t} D_{i,t'}, \end{aligned} \tag{3.15}$$

where the second equality holds because of the induction hypothesis. □

The alternative way of writing (3.10) is useful because it leads to a sparser set of equations that significantly improves the computational feasibility of the final model.

3.3.5 Cost factors

There are four cost factors in our model. Cost per labor hour during aggregated period $y \in Y$ is denoted c_y^W. For rotables not yet in the field in the first period of the planning horizon, a turn-around stock of rotables needs to be acquired at the price of c_i^a per rotable of type $i \in I$. (c_i^a may also include the expected inventory holding cost over the relevant time horizon.) There are also material costs associated with overhaul and these are denoted $c_{i,t}^m$ for rotables of type $i \in I$ when the overhaul order was released during period $t \in T_i^I$. Similarly, $c_{i,t}^r$ represent costs of replacing a rotable of type $i \in I$ during period $t \in T_i^I$. Note that we do not explicitly model the cost of replacing a rotable earlier than required; these costs can be modeled implicitly

through the dependence on time included in all the cost factors. Adding all costs over the relevant horizon, we find that the total relevant costs (TRC) satisfy

$$TRC = \sum_{y \in Y} c_y^W W_y + \sum_{i \in I | a_i > 1} c_i^a S_i + \sum_{i \in I} \sum_{t \in T_i^I} c_{i,t}^m n_{i,t} + \sum_{i \in I} \sum_{t \in T_i^I} c_{i,t}^r x_{i,t}. \tag{3.16}$$

3.3.6 Model remarks

In the above sections, we have given a mathematical description of the planning problem. In this description there are some implicit assumptions that we will highlight and justify in this section.

We assume many parameters to be deterministic and known, when in fact they are either random variables whose exact value will only become known later. Consider for example p_i, the period in which type i rotables become obsolete. Period p_i coincides with the end-of-life of the asset in which rotable i occurs. Companies plan the end-of-life of their assets, and so at least estimates of p_i are available in practical situations. We also note that these estimates typically become more accurate as the end-of-life of an asset becomes more imminent. Similar arguments can be made for a_i when $i \in I \setminus I_1$, r_i, q_i $\Delta_y^l(\Delta_y^u)$ etc. In fact for the decisions that are actually made based on the model, these parameters are known deterministically. Later into the future, the values of these parameters becomes increasingly uncertain. Note however that the decisions for these periods will not be implemented until the model is rerun after these parameters become known deterministically. These parameters and decision epochs are only included in the model to account for costs that are affected by current decisions but will occur much later in the life cycle of the involved asset.

Furthermore, we note that our model can easily deal with non-stationarity in the input over time, while stochastic models, generally cannot. As a final argument, we would like to point out that Dzielinski et al. (1963) and Spitter (2005) (Chapter 6) have tested deterministic rolling horizon models via simulation in dynamic and/or stochastic environments and have shown that these models perform favorably, and approach the performance of optimization models that do incorporate stochasticity. Comparisons with stochastic optimization models however, are only possible in relatively simple environments. In particular, it is usually assumed that stochastic quantities have stationary distributions over time, which, in our setting, is unlikely at best.

Table 3.2 Overview of notation.

		Sets
I	:	Set of all types of rotables (not the rotables themselves)
I_t	:	Set of all types of rotables in the field in period $t \in T$, $I_t = \{i \in I \mid a_i \leq t \leq p_i\}$
T	:	Set of all periods considered in the planning horizon (typically months)
T_i^I	:	Set of periods during which rotable $i \in I$ is active in the field ($T_i^I = \{a_i, \ldots, p_i\}$)
Y	:	Set of aggregated periods (typically years)
T_y^Y	:	Set of periods that are contained in a certain aggregated period $y \in Y$
		Input Parameters
a_i	:	First period in the planning horizon in which rotables of type $i \in I$ are in the field
B_i^{d}	:	Number of ready-for-use rotables of type $i \in I$ available (on stock) at the beginning of period a_i
c_i^a	:	Acquisition cost of rotable $i \in I \setminus I_1$
$c_{i,t}^m$	:	Material costs associated with releasing an overhaul order for rotable type $i \in I$ in period $t \in T$
$c_{i,t}^r$	:	Costs of replacing a rotable $i \in I$ during period $t \in T_i^I$
c_y^W	:	Cost per labor hour during aggregated period $y \in Y$.
$D_{i,t}^{\mathrm{d}}$	:	Number of rotables of type $i \in I$ that require overhaul in or before period $t \in \{a_i, \ldots, a_i + q_i\}$
H_i^{d}	:	Number of *non*-ready-for-use rotables of type $i \in I$ on stock at the beginning of period a_i
L_i	:	Overhaul lead time (in periods) for rotables of type $i \in I$
$n_{i,t}^d$	:	Number of overhaul order releases of rotables of type $i \in I$ in period $t \in \{a_i - L_i, \ldots, a_i - 1\}$
p_i	:	Last period in which rotables of type $i \in I$ are in the field during the planning horizon ($p_i \in T$)
q_i	:	Inter-overhaul deadline for rotables of type $i \in I$
r_i	:	Amount of labor hours required to start overhaul of a type $i \in I$ rotable
U_i^{d}	:	Number of replacements of rotables of type i in excess of what is strictly necessary by period $a_i - 1$
W^{d}	:	Number of labor hours available in the first aggregate period
$\Delta_y^l(\Delta_y^u)$	:	Lower (upper) bound on the change in number of labor contracts from aggregated period y to $y+1$, $y \in \{1, \ldots, \lvert Y\rvert - 1\}$
$\delta_t^l(\delta_t^u)$	:	Lower (upper) bound on labor hours for rotable overhaul made available in period $t \in T$ expressed as a fraction of $W_y/\lvert T_y^Y\rvert$, for $t \in T$.
		(Auxiliary) variables
$B_{i,t}$	:	Number of ready-for-use rotables of type $i \in I$ available at the beginning of period $t \in T_i^I$
$D_{i,t}$	:	Number of rotables of type $i \in I$ that require overhaul in or before period $t \in T_i^I$
$H_{i,t}$	:	Number of *non*-ready-for-use rotables of type $i \in I$ at the beginning of period $t \in T_i^I$
$U_{i,t}$	:	Number of replacements of rotables of type i in excess of what is strictly necessary by period t, i.e. $U_{i,t} = \sum_{t'=a_i}^{t} x_{i,t'} - \sum_{t'=a_i}^{t} D_{i,t'}$
		Decision variables
$n_{i,t}$	:	Number of overhaul order releases of rotables of type $i \in I$ during period $t \in \{a_i - L_i + 1, \ldots, p_i\}$
S_i	:	Turn-around stock of rotables of type $i \in I$
W_y	:	Number of labor hours available in aggregated period $y \in Y$
w_t	:	Number of labor hours for overhaul that are allocated to period $t \in T$
$x_{i,t}$	:	Number of rotable replacements of type $i \in I$ during period $t \in T$

3.3.7 Mixed integer programming formulation

The modeling results of the previous subsections lead to an optimization problem that we shall call the aggregate rotable overhaul and supply chain planning (AROSCP) problem. For convenience, all introduced notation is summarized in Table 3.2, where also a distinction is made between sets, parameters, (auxiliary) variables and decision variables. A natural formulation of AROSCP is a mixed integer program, as shown below.

$$\text{minimize} \quad TRC = \sum_{y \in Y} c_y^W W_y + \sum_{i \in I | a_i > 1} c_i^a S_i + \sum_{i \in I} \sum_{t \in T_i^I} c_{i,t}^m n_{i,t} + \sum_{i \in I} \sum_{t \in T_i^I} c_{i,t}^r x_{i,t} \tag{3.17}$$

subject to

$$B_{i,t} = B_{i,t-1} - x_{i,t-1} + n_{i,t-L_i-1} \qquad \forall i \in I, \quad \forall t \in T_i^I \backslash \{a_i\} \tag{3.18}$$

$$H_{i,t} = H_{i,t-1} + x_{i,t-1} - n_{i,t-1} \qquad \forall i \in I, \quad \forall t \in T_i^I \backslash \{a_i\} \tag{3.19}$$

$$B_{i,a_i} = S_i \qquad \forall i \in I \setminus I_1 \tag{3.20}$$

$$B_{i,a_i} = B_i^{\mathrm{d}} \qquad \forall i \in I_1 \tag{3.21}$$

$$H_{i,a_i} = 0 \qquad \forall i \in I \setminus I_1 \tag{3.22}$$

$$H_{i,a_i} = H_i^{\mathrm{d}} \qquad \forall i \in I_1 \tag{3.23}$$

$$n_{i,t} = n_{i,t}^{\mathrm{d}} \qquad \forall i \in I, \quad t \in \{a_i - L_i, \ldots, a_i - 1\} \tag{3.24}$$

$$W_y = \sum_{t \in T_y^Y} w_t \qquad \forall y \in Y \tag{3.25}$$

$$\delta_t^l W_y / |T_y^Y| \leq w_t \leq \delta_t^u W_y / |T_y^Y| \qquad \forall y \in Y, \quad \forall t \in T_y^Y \tag{3.26}$$

$$\Delta_y^l W_y \leq W_{y+1} \leq \Delta_y^u W_y \qquad \forall y \in \{1, \ldots, |Y| - 1\} \tag{3.27}$$

$$W_1 = W^d \tag{3.28}$$

$$\textstyle\sum_{i \in I_t} r_i n_{i,t} \leq w_t \qquad \forall t \in T \tag{3.29}$$

$$n_{i,t} \leq H_{i,t} + x_{i,t} \qquad \forall i \in I, \quad \forall t \in T_i^I \tag{3.30}$$

$$x_{i,t} \leq B_{i,t} + n_{i,t-L_i} \qquad \forall i \in I, \quad \forall t \in T_i^I \tag{3.31}$$

$$x_{i,t} \geq D_{i,t} - U_{i,t-1} \qquad \forall i \in I, \quad \forall t \in T_i^I \tag{3.32}$$

$$U_{i,t} = x_{i,t} - D_{i,t} + U_{i,t-1} \qquad \forall i \in I, \quad \forall t \in T_i^I \tag{3.33}$$

$$U_{i,a_i-1} = U_i^{\mathrm{d}} \qquad \forall i \in I \tag{3.34}$$

$$D_{i,t} = D_{i,t}^{\mathrm{d}} \qquad \forall i \in I, \quad \forall t \in \{a_i, \ldots, \min\{a_i + q_i - 1, p_i\}\} \tag{3.35}$$

$$D_{i,t} = x_{i,t-q_i} \qquad \forall i \in I, \quad \forall t \in \{a_i + q_i, \ldots, p_i\} \tag{3.36}$$

$$x_{i,t}, n_{i,t} \in \mathbb{N}_0 \qquad \forall i \in I, \quad \forall t \in T \tag{3.37}$$

$$S_i \in \mathbb{N} \qquad \forall i \in \{i \in I | a_i > 1\} \tag{3.38}$$

$$0 \leq n_{i,t}, x_{i,t}, B_{i,t}, H_{i,t}, U_{i,t} \qquad \forall i \in I, \quad \forall t \in T \tag{3.39}$$

$$0 \leq W_y \qquad \forall y \in Y \tag{3.40}$$

$$0 \leq w_t \qquad \forall t \in T. \tag{3.41}$$

Here, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. We remark that it is possible to choose parameter values such that a feasible solution to this MIP does not exist. In particular, infeasibility can be created by setting the parameters $D^{\mathrm{d}}_{i,t}$ to exceed the available capacity in terms of either workforce or rotable availability.

Because MIPs are hard to solve in general, it is natural to question what the complexity of AROSCP is. In this regard we offer the following proposition.

Proposition 3.2 *The aggregate rotable overhaul and supply chain planning problem (AROSCP) is strongly $\mathcal{NP}$-hard.*

The proof of Proposition 3.2 uses reduction from BIN-PACKING and is found in Appendix 3.A. In §3.4, we provide computational evidence that, despite the computational complexity of the problem, mixed integer programming can still be used to find optimal or close to optimal solutions for instances of real-life size.

3.3.8 Modeling flexibility

The formulation presented in (3.17)-(3.41) still has significant modeling flexibility. We shall illustrate this flexibility by several examples.

In many practical applications the availability of workforce fluctuates with the time of year; particularly during holiday and summer season there is reduced workforce availability. This can be modeled through the bounds on w_t, δ^u_t and δ^l_t.

The cost parameters in (3.17) depend on t. This dependence can be used to penalize early overhaul of rotables and to discount future costs, e.g. by taking $c^m_{i,t} = \alpha^t c^m_i$ with $\alpha \in (0, 1]$.

Labor flexibility has taken a very specific form that is congruent with the setting we will describe in §3.4. Traditionally, labor flexibility has been modeled by including overtime at extra cost in the model, as has also been done in Bitran and Hax (1977) and the related literature as reviewed in §3.2. These modeling constructs are easily incorporated into our MIP formulation.

In our model, we assume capacity bounds to exist only on labor in the overhaul workshop. The model can easily be extended with capacity constraints on the number

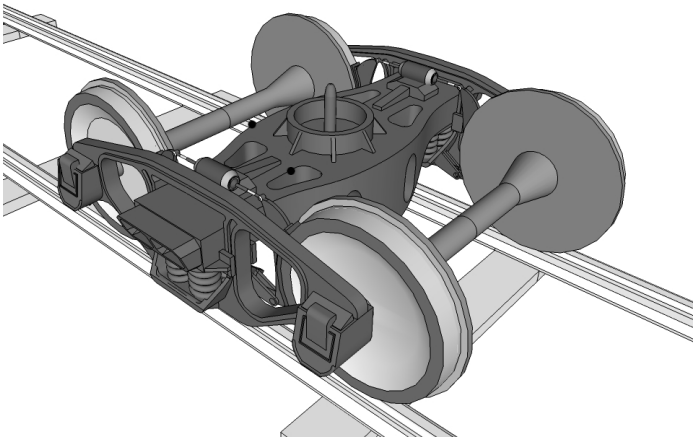

Figure 3.2 An example of a bogie.

of replacements in the maintenance depot and capacity constraints of different types (e.g. on equipment and tools) in both the overhaul workshop and the maintenance depot. Note however that when these constraints are clearly not binding, it is best to avoid the extra modeling and data collection efforts associated with such extensions.

3.4. Case study

In this section, we report on a case-study at NedTrain, a Dutch company that maintains rolling stock. The fleet maintained by NedTrain consists of some 3000 carriages across 6 main train types. Almost all carriages rest on two bogies. Bogies are rotables and there are about 30 different types of bogies in the fleet maintained by NedTrain. In the city of Haarlem, NedTrain has a facility dedicated to the overhaul of all types of bogies in the fleet. Bogies are considered important rotables and this case study is about the overhaul and supply chain planning of rotables at NedTrain. An example of a bogie is shown in Figure 3.2. The data set used for the case study we present is outlined in considerable detail in the master thesis of Vernooij (2011). Here, we present a high level description of the data. Rolling stock has a planned life cycle of 30 years and our model uses this as the length of the planning horizon. The period length we consider is a month, while the aggregated period length is a year. The instance we study has 56 bogie types, i.e. $|I| = 56$, 30 bogie types of which are currently in operation and 26 of which belong to new types of trains that will enter the fleet some time in the next 30 years. The population size of any bogie type ranges from 32 to 611 and depends on how many trains there are of a specific type in

the fleet, and how often a bogie type appears in any specific trainset. (For instance, bogies with traction engines appear less often than bogies without traction engines in most trainsets.) The flexibility of changing capacity from one aggregated period to the next is limited at 10%, i.e., $\Delta_y^l = 0.9$ and $\Delta_y^u = 1.1$ for all $y \in \{1, \ldots, |Y|-1\}$. The flexibility of allocating labor to specific periods is also limited to 10%, i.e., $\delta_t^l = 0.9$ and $\delta_t^u = 1.1$ for all $t \in T$. The MIOTs, q_i, range from 72 to 240 months. Overhaul lead times are 1 period for all bogie types ($L_i = 1$ for all $i \in I$). To start overhaul of any bogie, 200 hours of labor need to be available, for any bogie type ($r_i = 200$ for all $i \in I$). For confidentiality reasons, we do not report any real cost figures. Under the MIP formulation in this chapter, this instance has 64896 variables (42968 of which are auxiliary variables) and 76378 constraints.

3.4.1 Computational feasibility

Seeing as the AROSCP is $\mathcal{NP}$-hard, we first test the computational feasibility of the model. To this end we propose 3 ways to (approximately) solve the problem:

(i) Solve the MIP formulation while allowing for an optimality gap of 1%

(ii) Relax the integrality constraints on $n_{i,t}$ and $x_{i,t}$ and solve the resulting MIP while allowing for an optimality gap of 1%

(iii) Solve the linear programming relaxation of the MIP formulation.

All these three methods can be readily implemented using several MIP/LP solvers. We did this for four well known solvers: CPLEX 12.5.0.0[1] , GUROBI 4.6.1.[2] , CBC 2.7.5[3] , and GLPK 4.47[4]. We solved the instance of AROSCP described above 5 times for each combination of solver and (approximate) solution method. The average computational times and halfwidths of 95% confidence intervals based on the t-distribution are shown in Table 3.3. All experiments were ran on a machine with Intel Core Duo 2.93GHz processor and 4GB RAM. For the solvers, we used the 'out of the box' settings.

[1]CPLEX is a commercial solver that can use multiple CPU cores in parallel. More information on this solver can be found on http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

[2]GUROBI is a commercial solver that can use multiple CPU cores in parallel. More information on this solver can be found on http://www.gurobi.com/.

[3]CBC stands for Coin Branch and Cut and is an open source solver associated with the COIN-OR initiative. At present, CBC can only use one CPU core. More information on this solver can be found on http://www.coin-or.org/Cbc/.

[4]GLPK stands for GNU linear programming kit and is an open source solver. GLPK can only use one CPU core. More information on this solver can be found on http://www.gnu.org/software/glpk/.

Table 3.3 Computational times in seconds for different solvers and solution methods using 'out of the box' settings.

Solver	Solution Method	Average	Halfwidth of 95% CI
GUROBI 4.6.1	MIP (MIPGap 1%)	5128.0	27.83
	Partial MIP relaxation	119.2	0.30
	LP relaxation	85.6	0.60
CPLEX 12.5.0.0	MIP (MIPGap 1%)	out of memory after 881 s	
	Partial MIP relaxation	186.9	0.12
	LP relaxation	126.8	0.29
CBC 2.7.5	MIP (MIPGap 1%)	infeasible after 43200 s	
	Partial MIP relaxation	207.4	0.49
	LP relaxation	293.8	0.16
GLPK 4.47	MIP (MIPGap 1%)	infeasible after 43200 s	
	Partial MIP relaxation	3031.8	2.40
	LP relaxation	138.2	0.09

It is notable that only GUROBI can solve the MIP formulation; the other solvers either run out of memory or time. The reason for this seems to be that GUROBI generates more cuts, especially at the root node of the branch and bound tree. The result of this is that the branch and bound tree grows much slower compared to the other solvers. With a computational time of less than two hours, the performance of GUROBI is quite good. All solvers can solve the Partial MIP relaxation and the LP relaxation. The LP relaxation can be solved in a matter of minutes by any solver. In the next section, we show that the results produced by both the partial MIP relaxation and the LP relaxation are quite good in terms of both the estimated LCC and the decisions that follow from the solution.

3.4.2 Sensitivity of result to integrality constraints

The important tactical decisions that the model supports are the dimensioning of aggregate workforce capacity (W_y) and turn-around stocks (S_i). Figure 3.3 shows the aggregate capacity plan, W_y, for the planning horizon of 30 years as found by the three (approximate) solution methods proposed in §3.4.1. From Figure 3.3, it is evident that for tactical decision making, the results of both the Partial MIP relaxation and the LP-relaxation are sufficiently accurate, although the results of the Partial MIP relaxation are somewhat closer to the solution of the original MIP.

Results for the turn-around stock levels are also very close across different solution methods, as shown in Figure 3.4. Here the turn-around-stocks of the LP-relaxation

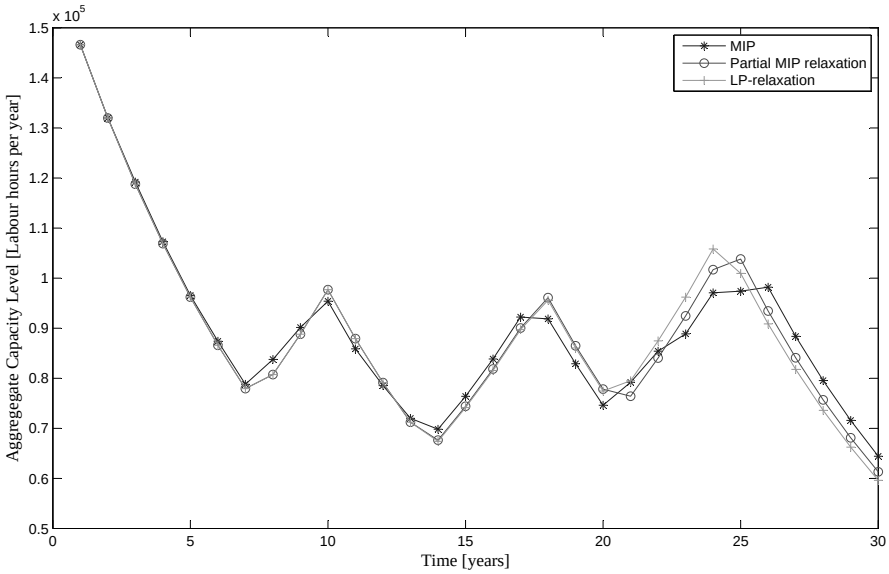

Figure 3.3 Aggregate capacity plan for case-study at NedTrain using different solution methods.

are determined by rounding up to the nearest integer. We remark that rounding the turn-around stock levels found by the LP-relaxation yields results that are closer to the MIP solution than the Partial MIP solution that does not relax integrality constraints on the turn-around stocks, S_i.

Figure 3.5 shows the costs found by all three solution methods, normalized so that the solution found by the MIP formulation is exactly 100. It is notable that estimated lower bounds of TRC found by solving either relaxation are very tight. Also Figure 3.5 shows that the division of costs over labor, material, acquisition and replacement costs are almost identical across solution methods, suggesting that the solution of the LP-relaxation does not only provide a tight lower bound, but also a similar solution that allocates costly resources in a similar manner. In conclusion, we observe that for making good decisions and estimating costs accurately, it is sufficient to solve relaxations of AROSCP. In particular the linear programming relaxation is a good candidate given its computational tractability.

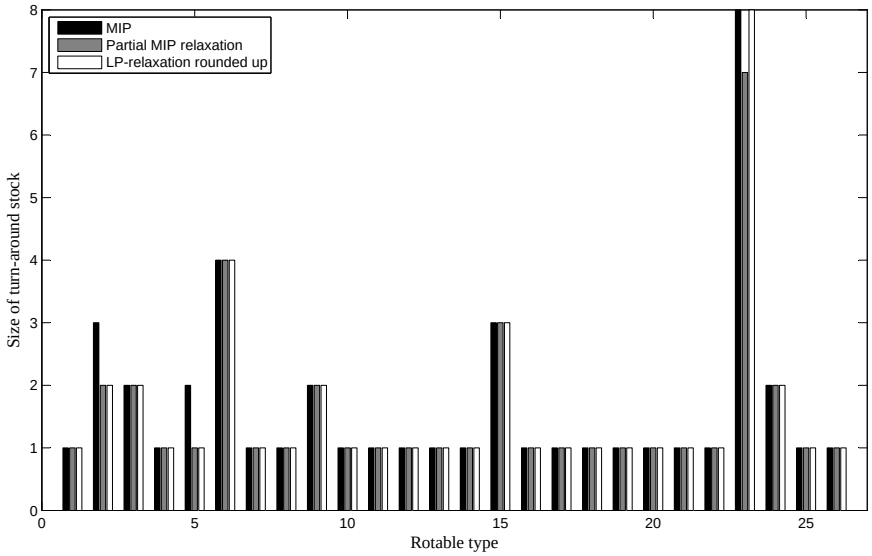

Figure 3.4 Turn-around stock sizes as determined by different solution methods.

3.4.3 Insights from case-study

From Figure 3.5, we know that labor costs are the most dominant cost factor. Our model allows for labor flexibility through the parameters Δ_y^u, Δ_y^l and δ_t^u, δ_t^l. The first two of these parameters control what we call long term labor flexibility, as they model how the size of the workforce can be changed over a longer horizon. The second pair of parameters, δ_t^u, δ_t^l, models the flexibility to allocate labor of the current workforce to different periods within the same aggregated period. For this reason, we say that δ_t^u, δ_t^l model short term labor flexibility. We performed a sensitivity analysis on long term versus short term labor flexibility. In what follows, we say that long (short) term labor flexibility is $x\%$ when $\Delta_y^u = 1 + x/100$ and $\Delta_y^l = 1 - x/100$ ($\delta_t^u = 1 + x/100$ and $\delta_t^l = 1 - x/100$) for all $y \in Y$ ($t \in T$). Figure 3.6 shows how TRC varies with different percentages of long and short term labor flexibility. Here again, costs were normalized to 100 for the MIP solution of the original instance with 10% labor flexibility in both the short and long term. It appears that short term labor flexibility has relatively little effect on costs over the horizon under consideration, while long term labor flexibility can be leveraged quite effectively. An explanation for this is that the greatest gains in planning rotable overhaul supply chains are often achieved by moving replacements and overhauls more than a year backward in time. Thus, effective planning does not rely on moving labor capacity around in the short term. Rather, gains can be made by

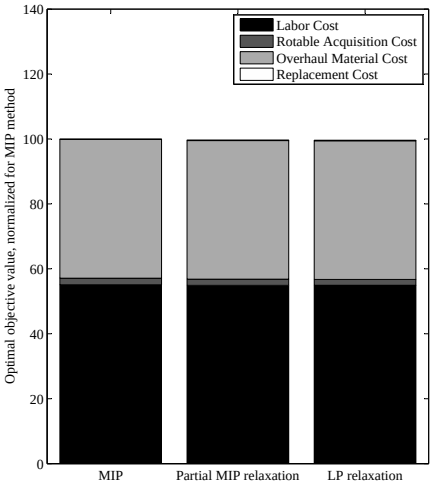

Figure 3.5 Cost break down for different solution methods.

planning replacement and overhauls such that exercising short term labor flexibility has only marginal impact. Overhauls and replacements interact with each other on the time scale of the MIOT. Thus, taking the entire life cycle of an asset and not artificially penalizing early overhaul and replacements really pays off.

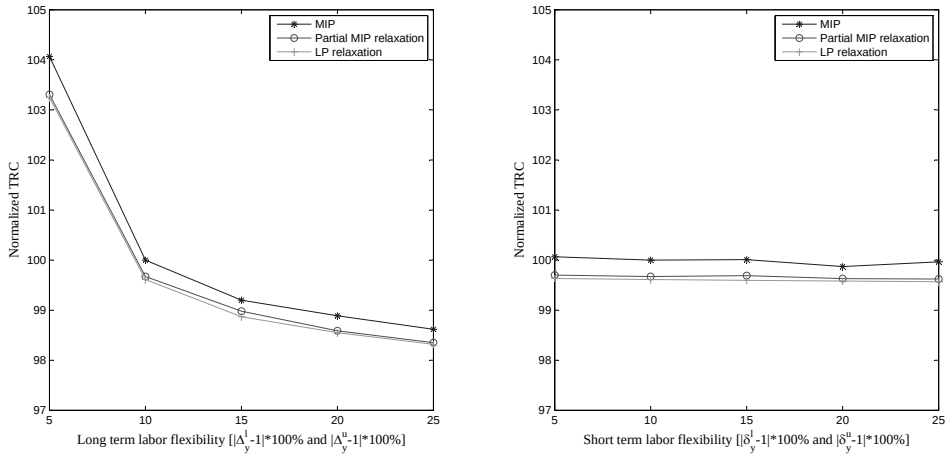

Figure 3.6 The value of long term versus short term labor flexibility

At NedTrain, it is practice not to plan overhauls and replacements very far into the future. The reason is that the MIOTs are subject to some uncertainty. The engineers that fix the MIOTs try to fix them as late as possible in the hope that they may stretch these MIOTs. The basic idea is that by stretching the MIOT, a rotable needs to undergo overhaul less often per time unit and so associated material and labor costs are incurred less often. While this is true for assets with an infinite life cycle and overhaul shops that have capacity available when convenient and not otherwise, it is not necessarily true for assets with a finite life cycle and overhaul shops that provide specialized labor that has to be contracted ahead of time. A result of knowing the MIOT late is that the overhaul workshop does not know how much work to expect, so it plans for the worst case scenario. Dealing with the worst case scenario requires keeping a large workforce for dealing with peak demand. If a large workforce is already in place, it is in fact optimal to postpone replacement and overhaul of rotables as long as possible. However, a better solution is to not focus on postponing overhaul, but on smoothing the workload of the overhaul shop because this is where most costs are incurred. Especially for the sake of labor costs and workload smoothing, it is much more beneficial to fix MIOTs early and optimize the plan for overhaul and supply chain operations. With low long term labor flexibility, the workforce remains dimensioned to deal with peak demand for overhaul capacity. Therefore, an indication of the possible cost savings of focussing on smoothing workload rather than postponing overhaul, can be read from Figure 3.6. If we choose to keep our workforce dimensioned for peak demand, costs increase by around 4%.

3.5. Numerical results for randomly generated instances

The results for the case study suggest that the LP-relaxation of our formulation yields sufficiently accurate results to aid decision making and that computation times are still practical. In this Section, we show that this behavior is typical by generating instances of the planning problem randomly and verifying that similar results are found. In §3.5.1, we give an overview of how instances are generated (pseudo) randomly and in §3.5.2, we explain the metrics we use to compare the LP-relaxation to the MIP optimum and discuss the numerical results.

3.5.1 Random instance generator

We generate instances randomly, but the orders of magnitude from which we generate values for these instances are based on the orders of magnitude observed at NedTrain, the company of the case study in §3.4. For a more detailed discussion of what these orders of magnitude are and how they arise, we refer to Vernooij (2011).

Table 3.4 shows how instances where generated (pseudo) randomly. A more detailed explanation of how instances are generated randomly is provided in Appendix 3.B. In Table 3.4 and Appendix 3.B, we use the notation $\mathcal{UD}(a,b)$ to denote a discrete uniform random variable on the integers $a,\cdots,b$ and $\mathcal{U}(a,b)$ to denote the (continuous) uniform random variable on the interval (a,b).

Table 3.4 Overview of how instances are generated randomly

Parameter	Generation	Index range
	Sets	
I_1	$\{1,2,3,\cdots,\mathcal{UD}(25,40)\}$	-
$I \setminus I_1$	$\{\lvert I_1\rvert+1,\cdots,\lvert I_1\rvert+\lvert\{j\in I_1 : p_j<360\}\rvert\}$	-
I	$I_1 \cup I \setminus I_1$	-
T	$\{1,2,\cdots,360\}$	-
Y	$\{1,2,\cdots,30\}$	-
T_y^Y	$\{12\cdot(y-1)+1,\cdots,12\cdot(y-1)+12\}$	$y\in Y$
	Rotable characteristics	
a_i	1	$i\in I_1$
p_i	$\min\{360,\mathcal{UD}(11,460)\}$	$i\in I_1$
q_i	$\mathcal{UD}(72,240)$	$i\in I_1$
r_i	$\mathcal{UD}(180,220)$	$i\in I_1$
L_i	1	$i\in I$
a_i	$p_{\min\{j\in I_1:\lvert\{k\in\{1,\cdots,j\}:p_k<360\}\rvert=i-\lvert I_1\rvert\}}+1$	$i\in I\setminus I_1$
p_i	360	$i\in I\setminus I_1$
q_i	$q_{\min\{j\in I_1:\lvert\{k\in\{1,\cdots,j\}:p_k<360\}\rvert=i-\lvert I_1\rvert\}}$	$i\in I\setminus I_1$
r_i	$r_{\min\{j\in I_1:\lvert\{k\in\{1,\cdots,j\}:p_k<360\}\rvert=i-\lvert I_1\rvert\}}$	$i\in I\setminus I_1$
	Initialization and flexibility	
τ_i	$a_i+\mathcal{UD}(10,q_i)$	$i\in I$
$D^{\mathrm{d}}_{i,\tau_i}$	$\mathcal{UD}(30,600)$	$i\in I$
$D^{\mathrm{d}}_{i,t}$	0	$i\in I, t\in\{a_i,\cdots,a_i+q_i\}\setminus\{a_i+\tau_i\}$
U^{d}_i	0	$i\in I$
$n^{\mathrm{d}}_{i,t}$	0	$i\in I, t\in\{a_i-L_i,\cdots,a_i-1\}$
H^{d}_i	0	$i\in I$
B^{d}_i	$\mathcal{U}(0.1,0.3)\cdot D^{\mathrm{d}}_{i,\tau_i}$	$i\in I$
W^{d}	150000	-
$\Delta^l(\Delta^u)$	$\mathcal{U}(0.7,0.95)\quad(\mathcal{U}(1.05,1.3))$	-
$\Delta^l_y(\Delta^u_y)$	$\Delta^l(\Delta^u)$	$y\in\{1,\cdots 29\}$
$\delta^l(\delta^u)$	$\mathcal{U}(0.7,0.95)\quad(\mathcal{U}(1.05,1.3))$	-
$\delta^l_t(\delta^u_t)$	$\delta^l(\delta^u)$	$t\in T$
	Cost parameters	
α	0.95	-
c^a_i	$\mathcal{UD}(300000,400000)\cdot\alpha^{a_i/12-1}$	$i\in I\setminus I_1$
$c^m_{i,t}$	$\mathcal{UD}(4000,6000)\cdot\alpha^{\lceil t/12\rceil-1}$	$i\in I,\quad t\in T^I_i$
$c^r_{i,t}$	$\mathcal{UD}(30,50)\cdot\alpha^{\lceil t/12\rceil-1}$	$i\in I,\quad t\in T^I_i$
c^W_y	$\mathcal{UD}(60,80)\cdot\alpha^{y-1}$	$y\in Y$

3.5.2 Results

In §3.4, the results of the original MIP and LP-relaxation are quite close, as evidenced by Figures 3.3-3.6. In the first experiment, we measure how 'close' the solutions of the MIP and LP-relaxation are by eight metrics. In this section, we use the superscripts LP and MIP on variables to denote that their values are obtained by solving the LP-relaxation and MIP formulation respectively. The eight metrics we consider are:

$$\Delta_{TRC} = \frac{TRC^{MIP} - TRC^{LP}}{TRC^{MIP}} \tag{3.42}$$

$$\Delta_W = \frac{\left|C_W^{LP} - C_W^{MIP}\right|}{C_W^{MIP}} \cdot 100\%, \qquad \text{with} \qquad C_W = \sum_{y \in Y} c_y^W W_y \tag{3.43}$$

$$\Delta_a = \frac{\left|C_a^{LP} - C_a^{MIP}\right|}{C_a^{MIP}} \cdot 100\%, \qquad \text{with} \qquad C_a = \sum_{i \in I: a_i > 1} c_i^a S_i \tag{3.44}$$

$$\Delta_m = \frac{\left|C_m^{LP} - C_m^{MIP}\right|}{C_m^{MIP}} \cdot 100\%, \qquad \text{with} \quad C_m = \sum_{i \in I} \sum_{t \in T_i^I} c_{i,t}^m n_{i,t} \tag{3.45}$$

$$\Delta_r = \frac{\left|C_r^{LP} - C_r^{MIP}\right|}{C_r^{MIP}} \cdot 100\%, \qquad \text{with} \quad C_r = \sum_{i \in I} \sum_{t \in T_i^I} c_{i,t}^r x_{i,t} \tag{3.46}$$

$$\Delta_{\text{capacity}}^{\max} = \max_{y \in Y} \left| \frac{W_y^{LP} - W_y^{MIP}}{W_y^{MIP}} \cdot 100\% \right| \tag{3.47}$$

$$\Delta_{\text{capacity}}^{\max(5)} = \max_{y \in \{1, \cdots, 5\}} \left| \frac{W_y^{LP} - W_y^{MIP}}{W_y^{MIP}} \cdot 100\% \right| \tag{3.48}$$

$$\Delta_S = \frac{\sum_{i \in I \setminus I_1} \left| \lceil S^{LP} \rceil - S^{MIP} \right|}{|I \setminus I_1|} \tag{3.49}$$

Metrics (3.42)-(3.46) measure the relative deviation of the objective function and the different terms of the objective function; together they convey the same information as Figure 3.5 does for the case study. Metrics (3.47)-(3.48) measure the relative deviation of aggregate capacity decisions, for the long term and the short term. In the case study, this information is conveyed by Figure 3.3. Finally, metric (3.49) measures the average absolute deviation of sizing the turn-around stock and conveys the information shown by Figure 3.4.

In §4.3, we already noted that not every instance of AROSCP is feasible. This is true in particular for instances generated randomly, as explained in §3.5.1. In this experiment, we generated instances until 200 feasible instances were obtained. To achieve this, a total of 280 instances were generated. For these 200 feasible instances, we solved the MIP formulation (while allowing for an optimality gap of 1%), and the LP-relaxation and computed metrics (3.42)-(3.49). Table 3.5 reports the results as well as the computation times (in minutes) on a machine with Dual Core 2.9 GHz processor with 4 GB of RAM and GUROBI 5.0 as solver.

Table 3.5 Accuracy of LP-relaxation in approximating an integer optimal solution. (N=200.)

	Δ_{TRC}	Δ_W	Δ_a	Δ_m	Δ_r	$\Delta^{\max}_{\text{capacity}}$	$\Delta^{\max(5)}_{\text{capacity}}$	Δ_S	CompTime [min]
avg	0.73	0.23	5.78	0.15	0.14	7.54	2.11	0.19	39.4
min	0.33	0.00	0.48	0.00	0.00	1.93	0.02	0.00	1.3
max	1.37	1.27	16.57	0.49	0.44	20.17	9.44	0.50	334.0
stdev	0.23	0.21	2.46	0.11	0.10	3.39	1.63	0.11	61.7

First, we note that the relative deviation with respect to the optimal objective value and its separate components is very small. Laying Δ_a aside for a moment, we see that Δ_{TRC}, Δ_W, Δ_m, Δ_r are all well below 1.0% on average and well below 1.5% in the worst case. This is remarkable, especially considering that the MIP solution still has a remaining optimality gap somewhere below 1.0%. The odd one out is Δ_a. We note however, that, as in the case study, the acquisition costs are a relatively small part of the total costs (as evidenced here by the fact that Δ_{TRC} is considerably smaller than Δ_a). Furthermore, the direct comparison of turn-around-stocks provided by the metric Δ_S is again quite favorable, probably also because S^{LP} is rounded up for the purpose of comparison. All in all, these results indicate that the LP-relaxation can be used to perform sensitivity analyses with respect to costs, thus saving considerable computation time. Furthermore, it is possible to use the dual variables provided by solving the LP-relaxation using the simplex method to streamline the sensitivity analysis. Consider, for example, the sensitivity of the model with respect to the bounds on capacity flexibility Δ^l_y and Δ^u_y. In the case study, the sensitivity to these bounds was investigated by repeatedly solving the problem, but via dual variables, it is possible to investigate the sensitivity of the optimal solution to these bounds around some operating point via dual-variables.

The measures, $\Delta^{\max}_{\text{capacity}}$ and $\Delta^{\max(5)}_{\text{capacity}}$ appear quite high. We note however that $\Delta^{\max}_{\text{capacity}} = 9.01\%$ and $\Delta^{\max(5)}_{\text{capacity}} = 0.35\%$ for the case study instance. Thus, performance is comparable to the instance in the case study.

In the second experiment, we investigate how computation time scales with the size of instances. The size of instances is controlled though the number rotables types that are in the field at any one time, $|I_1|$. We varied $|I_1|$ between 20 and 40 in steps

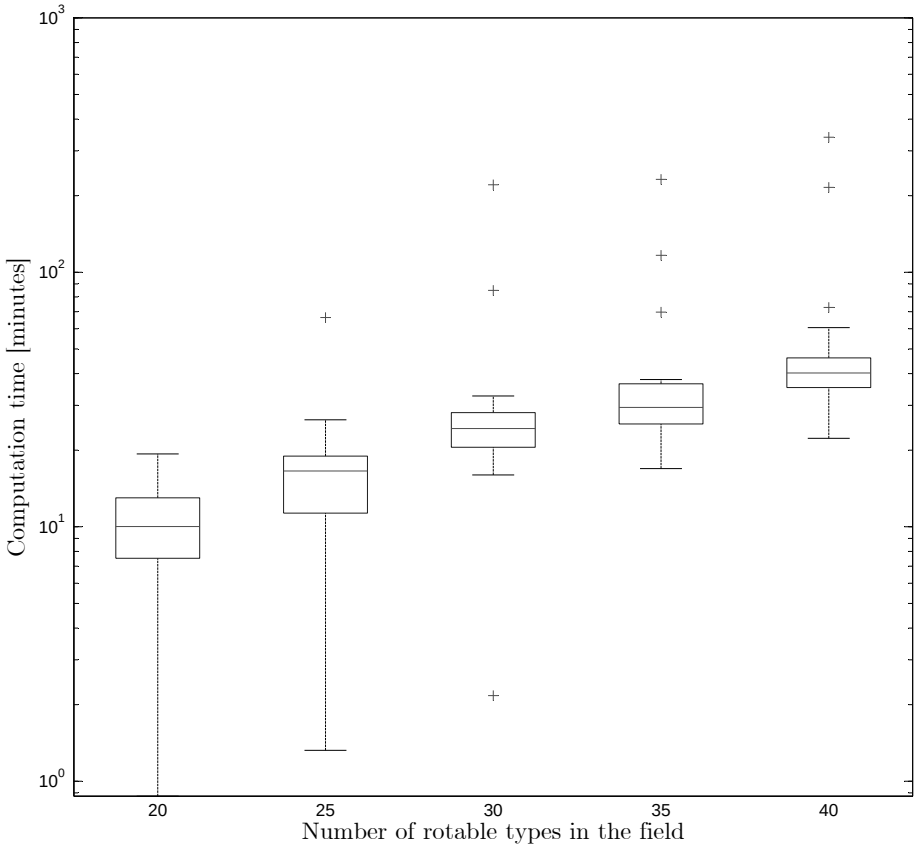

Figure 3.7 Boxplot of computation times for different instance sizes as measured by the number of rotable types in the field, $|I_1|$. For each of the sizes $|I_1| = 20, 25, 30, 35, 40$, 30 instances where generated randomly as shown in Table 3.4, with the exception that $|I_1|$ is controlled as indicated.

of 5 and generated 30 random instances for each size. Figure 3.7 shows a boxplot for $|I_1| = 20, 25, 30, 35, 40$. Note that the vertical axis has a logarithmic scale, so computation times seem to scale exponentially, which, given Proposition 3.2, is to be expected. There is also quite some variation in the running times for instances that are the same size. Nevertheless, the largest computation times for solving the largest instance are 5.7 hours, which is sufficiently short to do an overnight run. For sensitivity and scenario analyses, it seems the LP-relaxation provides a good

substitute with more acceptable computation times.

3.6. Conclusion

In this chapter, we have presented a model for the aggregate planning of rotable overhaul and supply chain operations. The model has many realistic features and incorporates LCC considerations in planning decisions when it is used in a rolling horizon setting. Despite the fact that solving the presented model to optimality is $\mathcal{NP}$-hard in general, we have provided evidence to suggest that a linear programming relaxation of the problem supplies the user with useful information that aids in decision making and even yields decisions that are close to optimal for large instances of AROSCP, as found in practice. In the context of a real life case study, we have argued that it is beneficial to fix MIOTs relatively early so that an effective plan for overhaul and supply chain operations can be made that utilizes the flexibility of overhaul planning that exists only when considering the entire life cycle of an asset. In the context of the case study, we also argued that it is better to focus on smoothing workload experienced by the overhaul workshop rather than focussing on postponing overhaul of rotables as long as possible.

3.A. Proof of Proposition 3.2

We show that being able to solve the AROSCP will enable one to decide the BIN-PACKING decision problem, i.e. we reduce BIN-PACKING to AROSCP. The following decision problem, known as BIN-PACKING, is strongly $\mathcal{NP}$-complete (e.g. Garey and Johnson, 1979): Given positive integers $\alpha_1, \ldots, \alpha_m$, β, and κ, is there a partition of $\{1, \ldots, m\}$ into disjoint sets $\Upsilon_1, \ldots, \Upsilon_\kappa$ such that $\sum_{j \in \Upsilon_i} \alpha_j \leq \beta$ for $i = 1, \ldots, \kappa$?

Now suppose we are given an instance of BIN-PACKING. Without loss of generality, we may assume that $\sum_{i=1}^{m} \alpha_i \leq \kappa\beta$, $\alpha_i \leq \beta$ for all $i \in \{1, \ldots, m\}$, and $\kappa \leq m$. From this instance of BIN-PACKING, we give a pseudo-polynomial transformation (as defined by (Garey and Johnson, 1979, p.101)) such that the answer to this instance of BIN-PACKING is yes, if and only if the optimal objective value of the corresponding instance of AROSCP equals 0. By Lemma 4.1 of Garey and Johnson (1979) we have then proven that AROSCP is strongly $\mathcal{NP}$-hard.

The basic idea behind this reduction is the following. By setting the initial number of non ready-for-use rotables sufficiently high, the release of overhaul orders is constrained only by available workforce capacity, i.e. by (3.29). This workforce capacity can be kept constant at β across periods by constraints (3.27) and (3.28). Now the problem can be viewed as packing overhaul order releases into several one period bins of fixed size β. By penalizing these order releases in all but κ periods, the objective becomes to pack as many order releases as possible in the κ periods in which the order releases are not penalized. If the optimal objective of AROSPC is 0, then it was possible to pack all overhaul order releases in κ periods and so the instance of BIN-PACKING is a yes-instance.

More formally, the reduction is as follows. Set $Y = \{1, \ldots, m+1\}$ and $T_y^Y = \{y\}$ for all $y \in Y$; thus, aggregated and regular periods coincide. Set $W^{\mathrm{d}} = \beta$, and $\Delta_y^l = \Delta_y^u = \delta_t^l = \delta_t^u = 1$. This ensures capacity is identical across periods. Set $I = \{1, \ldots, m\}$ and set $a_i = 1$, $p_i = m+1$, $q_i = m+1$, $L_i = 1$, $H_i^{\mathrm{d}} = 1$, $B_i^{\mathrm{d}} = 0$, $n_{i,0}^{\mathrm{d}} = 0$, $D_{i,m+1}^{\mathrm{d}} = 1$ and $r_i = \alpha_i$ for all $i \in I$. Furthermore, set $D_{i,t}^{\mathrm{d}} = 0$ for all $i \in I$ and $t \in \{1, \ldots, m\}$. Thus, each type of rotable needs to be replaced exactly once before or in the last period of the planning horizon. This instance of AROSCP is feasible because the following is a feasible solution: $n_{i,i} = 1$ for $i \in I$ and $n_{i,t} = 0$ otherwise, $x_{i,m+1} = 1$ for all $i \in I$ and $x_{i,t} = 0$ otherwise. (Note that all other variables are set by constraints). There are no acquisitions ($a_i = 1$ for all $i \in I$) so c_i^a does not need to be set. Most other cost parameters are set to 0; $c_y^W = 0$ for all $y \in Y$ and $c_{i,t}^r = 0$ for all $i \in I$ and $t \in T_i^I$. However, we set $c_{i,t}^m = 1$ for all $i \in I$ and $t \in \{1, \ldots, m-\kappa\}$ and set $c_{i,t}^m = 0$ otherwise. Note that $m - \kappa \geq 1$ because, by assumption, $\sum_{i=1}^{m} \alpha_i \leq \kappa\beta$ and $\alpha_i \leq \beta$ for all $i \in \{1, \ldots, m\}$. The objective

function now reduces to $\sum_{i\in I}\sum_{t=1}^{m-\kappa} c_{i,t}^m n_{i,t}$. Let OPT denote the optimal solution to this instance of AROSCP. If $OPT = 0$ then, necessarily $n_{i,t} = 0$ for all $i \in I$ and $t \in \{1,\ldots,m-\kappa\}$. Furthermore, by constraint (3.25), $\sum_{i\in I_t} r_i n_{i,t} \leq w_t$ for all $t \in T$, which, by our choice of parameter values, implies $\sum_{i\in I} \alpha_i n_{i,t} \leq \beta$ for all $t \in \{m-\kappa+1,\ldots,m\}$. All rotables in this instance of AROSCP have to be overhauled exactly once in or before period m because of constraints (3.31), (3.32) and (3.35). Therefore, for each $i \in I$, there is some $t \in \{m-\kappa+1,\ldots,m\}$ such that $n_{i,t} = 1$. when $OPT = 0$. Now it follows that a partition that satisfies the requirement of the original BIN-PACKING problem is given by:

$$\Upsilon_j = \{i \in I | n_{i,m-\kappa+j} = 1\}, \quad j \in \{1,\ldots,\kappa\}.$$

In an analogous manner, it is possible to construct an optimal solution with objective 0 to an instance of AROSCP if the corresponding instance and truth certificate of BIN-PACKING is given, by setting all $x_{i,t}$ and $n_{i,t}$ to 0, except $x_{i,m+1} = 1$ for all $i \in I$ and $n_{i,m-\kappa+j} = 1$ if $i \in \Upsilon_j$. Thus, we have shown that an instance of BIN-PACKING is a yes-instance if and only if the corresponding AROSCP problem has an optimal objective of 0. We observe further that (i) the reduction can be performed in polynomial time, (ii) that the numbers in the corresponding AROSCP instance are polynomially related to the numbers in the BIN-PACKING instance (in fact they are linearly related), and (iii) that the size of the AROSCP instance is polynomially related to the size of the BIN-PACKING instance (because $\kappa \leq m$). □

3.B. Details on the random instance generator

3.B.1 Rotable characteristics

First, we generate the number of different rotable types that are already in the field at the beginning of the planning horizon, $|I_1|$, from $\mathcal{UD}(25, 40)$. Then for each $i \in I_1$, we draw p_i, q_i, r_i and L_i as shown in Table 3.4. Note that for $i \in I_1$, $a_i = 1$ by definition and needs not be generated randomly.

Some of the rotables $i \in I_1$ may belong to assets that will be disposed of before the end of the planning horizon, i.e., possibly $p_i < 360$ for some $i \in I_1$. If this is the case, we assume that this asset will be replaced by a new type of asset, which consists of rotables with identical characteristics that will remain current for the remainder of the planning horizon. For example, if $1 \in I_1$, $p_1 = 270$, $a_1 = 1$, $q_1 = 120$, $r_1 = 200$, $L_1 = 1$, and $|I_1| = 32$, then we add rotable type 33 to I and set $a_{33} = p_1 + 1 = 271$, $p_{33} = p_1 = 360$, $q_{33} = q_1 = 120$, $r_{33} = r_1 = 200$, and $L_{33} = L_1 = 1$. This procedure is shown formally in Table 3.4 using set expressions and the fact that I is generated to contain a sequence of integers. We note that it is also possible that a rotable type that

is replaced some time during the planning horizon is replaced with a rotable type that has different characteristics. In particular, new rotable types are likely to be more reliable due to technological advancements. The models can also accommodate these scenarios. However, we make the conservative assumption that rotable types are replaced by rotable types with identical characteristics.

3.B.2 Initial conditions and flexibility

For each type of rotable $i \in I$, there are revisions already due. We assume the worst case scenario that the first upcoming revision of any one rotable type are due in a single period. For rotable type $i \in I$, this single period is $a_i + \tau_i$ and τ_i is generated as $\tau_i = \mathcal{UD}(10, q_i)$. The number of revisions due in period $a_i + \tau_i$ is drawn from $\mathcal{UD}(30, 600)$. This means that for each $i \in I$, $D^{\mathrm{d}}_{i,t} = 0$ if $t \neq \tau_i$ and $D^{\mathrm{d}}_{i,\tau_i} = \mathcal{UD}(30, 600)$.

Again as a worst case scenario, $U^{\mathrm{d}}_i = 0$ and $n^{\mathrm{d}}_{i,t} = 0$ for all relevant i and t, meaning that there are no recent order releases and that there have been no replacements ahead of time. Initially, for each $i \in I_1$ there is no stock of non-ready-for-use rotables, i.e., $H^{\mathrm{d}}_i = 0$ for all $i \in I_1$. The initial ready-for-use stock of $i \in I_1$ is generated as a fraction of the first peak number of revisions due in period $a_i + \tau_i$: $B^{\mathrm{d}}_i = \lceil \mathcal{U}(0.1, 0.3) \cdot D^{\mathrm{d}}_{i,\tau_i} \rceil$, where $\lceil x \rceil$ is the smallest integer equal to or exceeding x.

The bounds Δ^l_y and Δ^u_y are obtained by generating Δ^l (Δ^u) as $\mathcal{U}(0.7, 0.95)$ ($\mathcal{U}(1.05, 1.3)$) and setting $\Delta^l_y = \Delta^l$ and $\Delta^u_y = \Delta^u_y$ for all $y \in Y \setminus \{|Y|\}$. Similarly δ^l_t (Δ^u_t) are obtained by generating δ^l (δ^u) as $\mathcal{U}(0.7, 0.95)$ ($\mathcal{U}(1.05, 1.3)$) and setting $\delta^l_t = \delta^l$ and $\delta^u_t = \delta^u$ for all $t \in T$. Finally, in each case, W^{d} is set as 150000. We do not generate this parameter randomly, because the r_i are already generated randomly.

3.B.3 Costs parameters

The cost parameters are discounted on a yearly basis by the discount parameter α. Within a year however, there is no discounting, so that for a period $t \in T$, the corresponding year is $\lceil t/12 \rceil$.

Chapter 4

Optimal and heuristic repairable stocking and expediting in a fluctuating demand environment

"There can be only academic value in an 'optimal' policy"

Craig Sherbrooke

4.1. Introduction

Both service and manufacturing industries depend on the availability of expensive equipment to deliver their products. Examples of such equipment include aircraft, rolling stock and manufacturing equipment. When this equipment is not working, the primary processes of their owners come to an immediate stop. To reduce the downtime of equipment, companies stock critical components such that the equipment can be returned to an operational state quickly by replacing a defective component with a ready-for-use component. Many components represent a significant financial investment and so they are repaired rather than discarded after a defect occurs. Consider for example, aircraft engines, bogies, or lenses for wafer steppers; these are components of aircraft, rolling stock, and integrated circuit manufacturing equipment, respectively, and their prices range from several hundreds of thousands up to tens of millions of dollars. These expensive components are very specific to the equipment

they service. Consequently, the best time for companies to buy these components is at the same time as when the original equipment is purchased because, at this time, it is possible to negotiate reasonable prices. (In literature, this is often referred to as the initial spare parts supply problem and it occurs in many different environments (e.g. Rustenburg et al., 2001; Pérès and Grenouilleau, 2002). Later in time, such components often have to be custom made and prices are very steep if the component can be purchased at all. An aggravating factor is that demand intensity for these components typically fluctuates over time, reflecting the fluctuating need for maintenance over time. Companies anticipate these demand fluctuations by leveraging the possibility of expediting the repair of defective components, rather than buying new components. Expediting repair comes at a price, either because an external repair shop charges more for expedited repairs or because an internal repair shop can only handle a limited amount of expedited repairs. In the latter case, the cost of expediting can be thought of as a Lagrange multiplier that enforces a constraint on the number of expedited repairs that can be requested per time unit. In this situation, the model in the present chapter serves as a building block for a multi-item model. Chapter 5 explores such a multi-item model in detail.

In this chapter, we study a model that is inspired by practice at NedTrain. However, the problem is generic for companies that maintain capital assets and use repairable spare parts to run their operations such as in the aviation, defense, public transportation, and manufacturing industry. These companies hold inventory of repairable spare parts that are used to maintain equipment. When defective parts are sent to a repair shop for repair, there is often the possibility to request that the repair of a part is expedited. In this situation, these companies face two major decisions related to their inventory control, one at a tactical level, and another at the operational level:

1. How many repairable spare parts should we buy? (tactical)
2. When should we request that the repair of a part is expedited? (operational)

We refer to the first decision as the dimensioning decision and to the second as the expediting decision. The spare repairables are usually purchased at the same time as the technical systems which they support. After this time, the repairables are either no longer available in the market or prohibitively expensive. Thus the decision to buy repairables is a tactical decision that occurs one time only. The S spare repairables that are purchased at the time of the acquisition of the technical system are also called the *turn-around stock*. After this initial tactical decision come the operational recurring decisions to either expedite or not expedite the repair of spare parts each time demand occurs. These decisions should take demand fluctuations as well as current inventory levels into account. Expediting repairs incurs costs,

either because an external repair shop charges extra for expedited repairs, or because an internal repair shop needs to somehow adapt their operations to accommodate expedited repairs. The model in this chapter is intended to aid both the dimensioning and the expediting decision. Especially for the dimensioning decision, it is important to consider the fact that expediting will occur later at the operational level.

We study the decision problem of the previous paragraph via a stochastic inventory model for repairable items. In this model a defective item is replaced with a ready-for-use item and sent to a repair shop immediately after the defect occurs. At this point in time, the inventory manager is faced with the decision to either expedite or not expedite the repair of the part. Expediting repair is more costly but has a shorter lead time. This expediting decision is informed by knowledge about the fluctuation of demand intensity over time.

Our model runs in continuous time, and demand for the component is a Markov modulated Poisson process (MMPP). The state of the Markov chain that drives the demand process can be observed directly and is used to inform the expediting decision. This demand model is quite rich and can serve to model such divers things as economic conditions, seasons of a year, the degradation of a fleet of equipment, and knowledge about the maintenance program of equipment (Song and Zipkin, 1993). It has also been observed empirically that demand for repairable spare parts behaves as a non-stationary Poisson process (Slay and Sherbrooke, 1988). In any case, the MMPP offers the flexibility to model both stationary and non-stationary demand processes and so it can be used to model a wide variety of demand models. In particular it offers the possibility to model demand fluctuations.

We assume that there are no economies of scale in replenishment so that inventory is replenished by an $(S-1, S)$-policy, meaning that each defective item is immediately sent to the repair shop. We model the expedited lead time as being deterministic and the regular lead time as being the convolution of the expedited lead time and several exponential phases, the passing of which is monitored. Modeling lead time as such is a convenient device to investigate the value of tracking order progress information and the effect of different lead time distributions. (Gaukler et al., 2008, use a very similar model of order progress information.) Many lead time distributions can be modeled quite closely by this device and in particular deterministic lead times can be approximated arbitrarily closely by letting the number of exponential phases approach infinity.

The main contributions of this chapter are as follows. For the described setting, we characterize the optimal repair expediting policy for the infinite horizon average and discounted cost criteria by formulating the problem as a Markov decision process. We find that the optimal policy may take two forms. The first form is simply to never expedite repair. The second form is a state dependent threshold policy, where

the threshold depends on both the state of the modulating chain of demand and the pipeline of repair orders. We also provide monotonicity results for the threshold as a function of the pipeline of repair orders. We give closed-form conditions that determine which of the two forms is optimal. In analyzing the optimal policy, we also confirm a conjecture of Song and Zipkin (2009) that the expediting policy they propose is optimal for some special cases.

Secondly, we show how to optimally solve the joint problem of determining the turn-around stock and the expediting policy.

Thirdly, we propose a heuristic that is computationally efficient, and is shown to perform well compared to the optimal solution. In this heuristic, we replace the optimal expediting policy with a parameterized threshold policy that shares important monotony properties with the optimal expediting policy. The thresholds depend on the available knowledge about the fluctuation of demand. Borrowing the terminology of Zipkin (2000), we call this policy the world driven threshold (WDT) policy. In a numerical study involving a large test bed, this heuristic has an average and maximum optimality gap of 0.15% and 0.76% respectively.

Finally, we investigate the value of anticipating demand fluctuations by comparing optimal joint stocking and expediting policy optimization against naive heuristics that do not explicitly model demand fluctuations, or that separate the stocking and expediting policy decisions. These naive heuristics have optimality gaps of 11.85% on average and range up to 63.67% in our numerical work. The comparison with these naive heuristics show that

1. There is great value in leveraging knowledge about demand fluctuations, in making repair expediting decisions.

2. Fluctuations of demand and the possibility to anticipate these through expediting repairs should be considered explicitly in sizing the turn-around stock and can lead to substantial savings.

This chapter is organized as follows. In §4.2, we review relevant literature and position our contribution with respect to existing results. The model is described in §4.3 and analyzed exactly in §4.4. In general, the exact analysis leads to algorithms that suffer from the curse of dimensionality. Therefore, in §4.5, we study a heuristic informed by our exact analysis that is computationally tractable. In §4.6, we provide numerical results on the performance of the heuristic we propose and investigate the value of anticipating demand fluctuations through the joint optimization of the turn-around stock and expediting policy. Concluding remarks are provided in §4.7.

4.2. Literature review

Our model is situated at the intersection of two streams of literature. The first one deals with sizing the turn-around stock of repairable item inventories and the second one with expediting, or inventory models with two (or more) supply modes.

An important characteristic of repairable item inventories is that inventory is replenished by repairing defective items. Repairable item inventory systems thus form a closed loop system that implicitly dictates base-stock levels. Often, the number of supported assets is large and the demand process is assumed to be independent of the number of outstanding orders. A small stream of literature considers situations where the number of supported technical systems is low and so the number of outstanding orders will affect the demand process (e.g. Gross and Ince, 1978). We assume that demand for the repairable is not affected by the number of outstanding repair orders. This is in line with the modeling assumptions of most of the repairable item inventory literature that was started with the METRIC model introduced by Sherbrooke (1968). Most of the important results in this stream of literature have been consolidated in the books by Sherbrooke (2004) and Muckstadt (2005). In this chapter, we add to the literature on repairable item inventories by studying what happens when it is possible to expedite the repair of a defect part, and in particular if this flexibility can be used to respond to a fluctuating demand environment. In doing this, we relax the commonly held assumption that demand is a stationary Poisson process. Our assumption of a Markov modulated Poisson process is more in line with empirical findings (Slay and Sherbrooke, 1988). Verrijdt et al. (1998) already studied simple heuristics for the case that demand is a stationary Poisson process and emergency and regular repair lead times are both exponentially distributed. We relax the assumptions that the demand process is stationary and consider a more general lead time structure. Furthermore, we study optimal solutions as well as a new heuristic informed by the structure of the optimal solution. We also remark that expediting repair is not the same as shipping a ready-for-use part from a different stocking location which is commonly known as an emergency shipment (e.g. Alfredsson and Verrijdt, 1999).

Inventory models with multiple supply modes have been reviewed by Minner (2003). Here we review the important and more recent results. Most authors consider a *periodic review* setting where the regular and expedited lead time differ by a single period and find that a base-stock policy is optimal for both the regular and expedited supply modes (e.g. Fukuda, 1964). When the lead time of the regular and expedited supply modes differ by more than a single period, optimal policies do not exhibit simple structure and depend on the entire vector of outstanding orders (e.g. Whittmore and Saunders, 1977; Feng et al., 2006). As a result, recent research considers heuristic policies for the control of dual supply systems, the most notable of

these being the dual-index policy and variations thereof (Veeraraghavan and Scheller-Wolf, 2008; Sheopuri et al., 2010; Arts et al., 2011). Under the dual-index policy, a regular and emergency inventory position are tracked separately, and both are kept at or above their respective order-up-to levels.

As opposed to the above mentioned papers, Moinzadeh and Schmidt (1991) consider a system running in *continuous time* facing Poisson demand with deterministic emergency and regular replenishment lead times. They show how to evaluate a given dual-index policy, although the name was not coined at the time, and the structure was not recognized as such. Song and Zipkin (2009) reinterpret the model of Moinzadeh and Schmidt (1991) revealing the simple structure of the policy and show how the performance of any such policy can be evaluated in closed form using an equivalence to a special type of queueing network that has a product form solution. This equivalence also allows them to consider very general lead time structures. Verrijdt et al. (1998) consider a similar system in the context of repairable items. In their model, the regular and expedited supply/repair modes have independent exponentially distributed lead times. They consider a different policy where repair is expedited when the inventory on hand drops below a certain critical level.

While two different heuristic expediting policies have been suggested in the literature by Moinzadeh and Schmidt (1991) and Song and Zipkin (2009), and Verrijdt et al. (1998), the optimal expediting policy has not yet been investigated. Song and Zipkin (2009) conjecture that their policy is optimal in some special cases. In this chapter, we analyze the optimal repair expediting policy in the case of deterministic expedited repair lead times and stochastic regular repair lead times. As it turns out, the form suggested by Moinzadeh and Schmidt (1991) and Song and Zipkin (2009) is optimal in the special case that the regular repair lead time has a shifted exponential distribution and demand is a Poisson process. For more general lead time structures and demand processes, the optimal policy is a generalization of this policy. We note that Song and Zipkin (2009) also considered Markov modulated Poisson demand as an extension, but their expediting policy does not depend on the state of modulating chain of demand.

4.3. Model formulation

Our model supports two decisions: (i) How to dimension the turn-around stock S and (ii) what expediting policy to follow. The two decisions we consider in this chapter live in different time scales. For the analysis, we will use a nested procedure that determines the optimal expediting policy for a given turn-around stock, and use this to determine the optimal turn-around stock. Below we give an integrated description of the model. In §4.3.1, we discuss the main assumptions of the model and their

justifications.

We consider a repairable item stock-point operated in continuous time with an infinite planning horizon $[0, \infty)$. The stock-point faces Markov modulated Poisson demand and so demand is a Poisson process whose intensity varies with the state of an exogenous Markov process $Y(t)$. The Markov process $Y(t)$ is irreducible and has a finite state space $\Theta = \{1, ..., |\Theta|\}$ with generator matrix $\mathbf{Q}$ whose elements we denote by q_{ij}. For notational convenience, we also define $q_i = -q_{ii}$ and $q_{\max} = \max_{i \in \Theta} q_i$. When $Y(t) = y$, the intensity of Poisson demand is given by $\lambda_y \geq 0$; $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_{|\Theta|})$, $\lambda_y > 0$ for at least one $y \in \Theta$. For convenience, we also define $\lambda_{\max} = \max_{y \in \Theta} \lambda_y$. We denote demand in the time interval $(t_1, t_2]$ given $Y(t_1) = y$ as $D^y_{t_1,t_2}$. Note that $Y(t_1)$ provides information about the distribution of demand in the interval $(t_1, t_2]$, $t_2 > t_1$. We assume that $Y(t)$ can be observed by the decision maker and so it provides a form of aggregated advance demand information.

The size of the turn-around stock, $S \in \mathbb{N}_0$, of the repairable is determined at time $t = 0$ and cannot be adapted afterwards. We assume that failed parts can always be repaired (no condemnation) and that defective parts are sent to the repair shop immediately, i.e., we use an $(S - 1, S)$ replenishment policy.

There exists a regular and an expedited repair option. The expedited repair lead time, ℓ_e, is deterministic. The expedited repair lead time may represent things such as the transport time and the repair time or a lead time agreed upon with an external company that provides emergency repair service. We also refer to using the expedited repair mode as expediting repair.

The regular repair lead time consists of the emergency repair lead time ℓ_e, and a random component of length L_r, with mean $\mathbb{E}[L_r] < \infty$. We shall also refer to L_r as the *additional regular repair lead time*. L_r is used to model such things as the time that a part waits for resources to become available in the repair shop or the lead time difference between regular and emergency repair lead times as contracted with an external repair shop. We assume that this additional time is distributed as the sum of m exponential phases, with mean $1/\mu_i$ for the i-th exponential phase. We also let $\mu_{\max} = \max_{i \in \{1,...,m\}} \mu_i$. The inventory manager can observe the pipeline of outstanding orders and so she knows how many phases each part in the pipeline has completed. In particular, the inventory manager knows when the last phase (m) is completed and the remaining lead time of a regular order is ℓ_e. A graphical representation of the system under study is given in Figure 4.1.

Turn-around stock depreciation costs are incurred with a constant rate $h > 0$ for all repairable spare parts, regardless of where they are in the supply chain. Repair expediting costs per item are $c_e > 0$, i.e., c_e represents the cost difference between using the regular and emergency repair modes. A penalty cost rate $p > 0$ per item

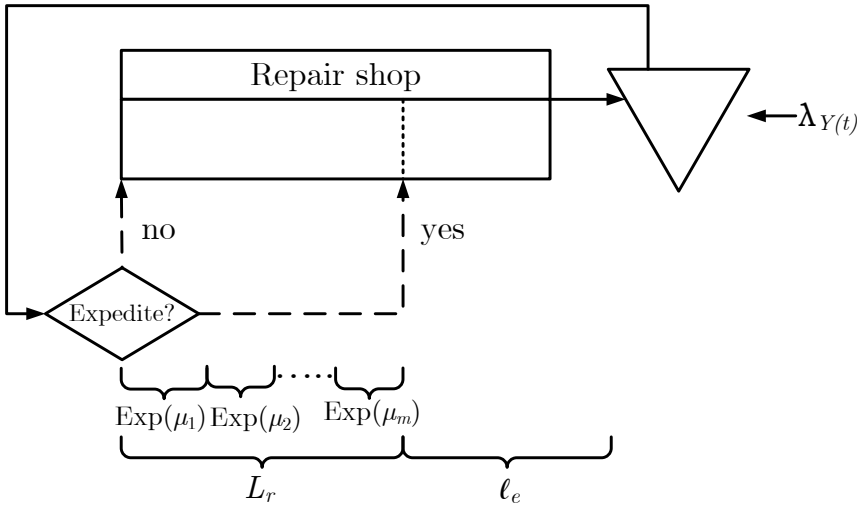

Figure 4.1 Repairable item inventory system with the possibility to expedite repair

short per time unit for the repairable item inventory is also charged (backordering). We are interested in minimizing the long run average cost rate by (i) deciding on the turn-around stock, S, to be purchased at time $t = 0$ and (ii) implementing a repair expediting policy. (As an extension, we shall also consider minimizing total discounted costs for the expediting policy in §4.4.1.4.)

4.3.1 Main assumptions and justifications

In the model of §4.3, some assumptions require either a practical or analytical justification. Here we list the main assumptions and their justifications.

- The turn-around stock S is determined at time $t = 0$, and remains fixed after that: Because repairables are specific to the capital asset which they support, they are only produced in small series when the capital asset is produced. After the particular capital asset is no longer produced, the repairable is either no longer available or has to be custom made against a steep price. Thus, for the user of the capital asset, it is most economical to purchase all spare repairables jointly with the asset they support.
- We consider an infinite planning horizon. The lifetime of repairables considered in the model is as long as the life cycle of the assets they support which is typically several decades. This is long compared to other time characteristics

in the problem such as lead times which are typically measured in weeks, and justifies using infinite horizon models.

- Demand is a Markov modulated Poisson process: In spare parts literature, the Poisson demand model is perhaps the most common (e.g. Sherbrooke (2004) and Muckstadt (2005)). For relatively short periods of time, this demand model is often sufficiently accurate, and Markov modulated Poisson demand can handle Poisson demand as a special case. For longer periods of time, the demand intensity for repairables may be affected by things such as weather conditions (increased wear) and periodic inspections. Slay and Sherbrooke (1988) observe that demand for aircraft components behaves as a Poisson process for which the rate varies over time. There are many reasons for this behavior such as weather, asset loading, and the fact that many capital assets undergo one or more major revisions during their lifetime. During these revision periods, demand for repairables peak, as inspections reveal latent failures. Often, the exact timing of revision periods is uncertain when the asset is acquired. The Markov modulated Poisson process offers the flexibility to model these and many other demand scenarios. The recent trends in condition based maintenance are a rich source for modeling demand using a MMPP.

- Repair of a part is always possible (no condemnation): Under normal operations, the expensive repairables considered in our model only fail permanently in case of industrial accidents. Generally, the probability of this happening is negligible.

- The additional regular repair lead time, L_r, can be modeled by a sum of exponential phases, and phase transitions can be observed. This assumption appears quite strong. We think of the m exponential phases of L_r primarily as a device to model order progress information. A special case occurs as $m \to \infty$ as this will approach deterministic replenishment lead times and order progress is known exactly. We also note that if the first two moments of L_r are known (or estimated from data) and satisfy $c^2_{L_r} = \mathbf{Var}[L_r]/\mathbb{E}^2[L_r] \leq 1$, then m and μ_i, $i = 1, ..., m$ can be chosen so as to match these moments. Such a fitting procedure will require that $m \geq \lfloor 1/c^2_{L_r} + 1 \rfloor$ (Aldous and Shepp, 1987). It is evident that as $c^2_{L_r}$ decreases, more information is available on when a repairable completes its additional regular repair lead time. Under the present model, this is naturally matched by increasing m. Therefore, we may think of the parameter m as a modeling device that conveys how closely one tracks, or is able to track, the progress of repairables through the replenishment pipeline. In particular, as $m \to \infty$, the regular replenishment lead time approaches a deterministic lead time and order progress is known exactly. Thus, this assumption allows us to gain insight on the added value of being able to track repairable order progress carefully. However in §6.8, we show that this added value is small and

consequently we believe it is unnecessary to refine the model of the additional regular repair lead time and how progress through the orde pipeline is, or can be, monitored.

4.4. Exact Analysis

The analysis of the model benefits from first considering the optimization of the expediting policy separately. That is, we use a nested procedure. Therefore in §4.4.1, we consider our model where the turn-around stock, S, is fixed, and focus on finding an optimal expediting policy. We call this problem $\mathfrak{M}(S)$. After this, we turn attention to the *joint* problem of sizing the turn-around stock and determining an expediting policy in §4.4.2.

4.4.1 Expediting policy optimization

In this subsection, we consider the problem of finding optimal repair expediting policies for fixed S. (Recall that this problem was termed $\mathfrak{M}(S)$). Since the holding costs depend linearly on S only, we need not consider holding cost in finding an optimal expediting policy for a fixed S. We make several steps in our analysis. First, we give the state space description and give closed form conditions under which the state space can be truncated to yield a finite state space for the purpose of finding average optimal expediting policies. We also show that when these conditions do not hold, the optimal policy is to never expedite repair. After that, we formulate a finite horizon finite state space Markov decision process. The average optimal expediting policy is characterized in §4.4.1.3 and the infinite horizon discounted version in §4.4.1.4.

4.4.1.1 State space description

Let $X_i(t)$ denote the number of items in regular repair at time t that are in the i-th phase of their additional repair lead time ($i = 1, ..., m$), and let $\mathbf{X}(t) = (X_1(t), ..., X_m(t))$. The following observation shows that $\mathbf{X}(t)$ and $Y(t)$ contain all the information needed to make expediting decisions. Let $c_p(x, y)$ denote the expected penalty cost rate at time $t+\ell_e$ conditional on $\sum_{i=1}^{m} X_i(t) = \mathbf{X}(t)\mathbf{e}^{\mathrm{T}} = x$ and $Y(t) = y$, $c_p : \mathbb{N}_0 \times \Theta \to \mathbb{R}$ ($\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and $\mathbf{e} = (1, 1, \ldots, 1)$). To find $c_p(x, y)$, note that $S - \mathbf{X}(t)\mathbf{e}^{\mathrm{T}} = S - x$ represents the number of parts that are on hand at the stockpoint at time t or will arrive at the stockpoint before time $t + \ell_e$. Thus the expected number of backorders at time $t + \ell_e$ given $\mathbf{X}(t)\mathbf{e}^{\mathrm{T}} = x$ and $Y(t) = y$ is $\mathbb{E}\left[\left(D_{t,t+\ell_e}^{Y(t)} - (S - \mathbf{X}(t)\mathbf{e}^{\mathrm{T}})\right)^{+} \mid \mathbf{X}(t)\mathbf{e}^{\mathrm{T}} = x, Y(t) = y\right]$. From this it is easily verified

that

$$c_p(x,y) = p\,\mathbb{E}\left[\left(D^{Y(t)}_{t,t+\ell_e} - \left(S - \mathbf{X}(t)\mathbf{e}^{\mathrm{T}}\right)\right)^+ \middle| \mathbf{X}(t)\mathbf{e}^{\mathrm{T}} = x, Y(t) = y\right]$$
$$= p\sum_{k=S-x}^{\infty}(k-(S-x))\,\mathbb{P}\left\{D^{y}_{t,t+\ell_e} = k\right\}. \tag{4.1}$$

When convenient, we also use the notation $c_p(x,y|S)$ for $c_p(x,y)$ to make the dependence on S explicit. We note that to use (4.1), one must be able to evaluate $\mathbb{P}\left\{D^{y}_{t,t+\ell_e} = k\right\}$. This can be done numerically by inverting the generating function of $\mathbb{P}\left\{D^{y}_{t,t+\ell_e} = k|Y(t+\ell_e) = y'\right\}$ which is given in the form of a matrix exponential (e.g. Fischer and Meier-Hellstern, 1992) and then un-conditioning on the event $Y(t+\ell_e) = y'$. We relegate further details of this to appendix 4.A. Let Δ denote the difference operator with respect to the first argument of a function, i.e., $\Delta c_p(x,y) = c_p(x+1,y) - c_p(x,y)$. The following lemma establishes some useful properties of c_p. The proof of Lemma 4.1, in Appendix 4.B.1, is similar to the proof of these same properties for the cost function of a news-vendor problem.

Lemma 4.1 *$c_p(x,y)$ has the following properties:*

(i) $c_p(x+1,y) \geq c_p(x,y)$ *for all* $x \in \mathbb{N}_0$ *and* $y \in \Theta$*, i.e., c_p is non-decreasing in x.*

(ii) $\Delta c_p(x+1,y) \geq \Delta c_p(x,y)$ *for all* $x \in \mathbb{N}_0$ *and* $y \in \Theta$*, i.e., c_p is convex in x.*

(iii) $\Delta c_p(x,y) \leq p$ *for all x and* $y \in \Theta$ *and* $\Delta c_p(x,y) = p$ *for all* $x \geq S$ *and* $y \in \Theta$*.*

(iv) $\Delta c_p(x,y|S) \geq \Delta c_p(x,y|S+1)$ *for all* $x \in \mathbb{N}_0$*,* $y \in \Theta$ *and* $S \in \mathbb{N}_0$*.*

Next, we note that whenever an item fails at time t, and is not expedited, $X_1(t)$ increases by one. Thus, $\mathbf{X}(t)$ and $Y(t)$ contain all information needed to do cost accounting, and, in particular, to make optimal expediting decisions.

Proposition 4.1 below allows us to truncate the relevant state space if $c_e < p\mathbb{E}[L_r]$, and fully characterizes an optimal expediting policy if $c_e \geq p\mathbb{E}[L_r]$. Proposition 4.1 can be understood intuitively by making the following casual observation: Whenever a repair is expedited, this may avert a backorder at most for the additional regular repair lead time L_r. Thus, when the cost of expediting repair is more than or equal to the expected backorder cost over the additional regular repair lead time, expediting is never beneficial. Conversely, if expediting is cheaper than the cost of a backorder over the expected additional regular lead time, then expediting is almost certainly beneficial if the number of parts already in repair that will not arrive within the expedited lead time is sufficiently large.

Proposition 4.1 *For the infinite horizon, average cost criterion, the following statements hold:*

(i) If $c_e \geq p\mathbb{E}[L_r]$ then it is optimal to never expedite repair.

(ii) If $c_e < p\mathbb{E}[L_r]$ then there is an $M \in \mathbb{N}$ such that whenever $\mathbf{X}(t)\mathbf{e}^T \geq M$ it is optimal to expedite repair upon failure of a part.

PROOF: Here we prove part (i). The proof of part (ii) is in the appendix; that proof is more subtle, involving the verification that several limits exist, but is based on a similar idea.

The proof is based on showing that any policy that expedites in some state when $c_e \geq p\mathbb{E}[L_r]$ can be improved by a policy that is identical except that it does not expedite in that state. Let π denote an arbitrary policy that expedites for some state $(\mathbf{x}, y)$. Suppose now that at time t', the process is in state $(\mathbf{x}, y)$ and a demand occurs. Let $(\mathbf{X}(t), Y(t))$ denote the process under policy π. Next we construct a coupled process, $(\mathbf{X}'(t), Y(t))$, that is identical to $(\mathbf{X}(t), Y(t))$ except that the failed part arriving at time t' is *not* expedited. Let $\widetilde{\mathbf{X}}(t)$ denote the evolution of the part expedited at time t' by policy π through the pipeline, i.e., $\widetilde{\mathbf{X}}(t) = \mathbf{e}_i$ if the part sent to regular repair at time t' has completed its first $i-1$ phases of the additional regular repair lead time at time t, and $\widetilde{\mathbf{X}}(t) = \mathbf{0}$ if the part has completed its additional regular repair lead time. ($\mathbf{e}_i$ is the i-th unit vector with dimension m.) With this notation, we can write $\mathbf{X}'(t) = \mathbf{X}(t) + \widetilde{\mathbf{X}}(t)$. Now let $T_r = \inf\{t - t' | \widetilde{\mathbf{X}}(t) = \mathbf{0},\ t \geq t'\}$ and note that $T_r \stackrel{d}{=} L_r$, where $\stackrel{d}{=}$ denotes equality in distribution. By construction, any cost difference between the processes $(\mathbf{X}'(t), Y(t))$ and $(\mathbf{X}(t), Y(t))$ must occur in the interval $[t', t' + T_r)$, because these processes are identical outside that interval. In $[t', t' + T_r)$, $\mathbf{X}(t)$ incurs exactly c_e more emergency repair costs due to the part expedited at time t', and $\mathbf{X}'(t)$ incurs more penalty costs because $\mathbf{X}'\mathbf{e}^{\mathrm{T}} = \mathbf{X}(t)\mathbf{e}^{\mathrm{T}} + 1$ for $t \in [t', t' + T_r)$. The expected cost difference between the processes $(\mathbf{X}(t), Y(t))$ and $(\mathbf{X}'(t), Y(t))$ thus satisfies:

$$c_e - \mathbb{E}_{T_r}\left\{\mathbb{E}_{(\mathbf{X}(t),Y(t))}\left[\int_{t=t'}^{t'+T_r} \Delta c_p(\mathbf{X}(t)\mathbf{e}^{\mathrm{T}}, Y(t))\middle| T_r\right]\right\} \geq c_e - \mathbb{E}_{T_r}[pT_r]$$
$$= c_e - p\mathbb{E}[L_r] \geq 0 \qquad (4.2)$$

where the first inequality holds by lemma 4.1 (iii). Thus we see that when $c_e \geq pE[L_r]$, any policy π that expedites for some states, can be improved (in the weak sense) by changing the decisions to not expedite in those states. This implies that when $c_e \geq p\mathbb{E}[L_r]$, the policy to never expedite is optimal. □

Proposition 4.1 has an important implication: When $c_e < p\mathbb{E}[L_r]$, there is a finite M such that it is optimal to expedite repair in all states $(\mathbf{x}, y)$ such that $\mathbf{x}\mathbf{e}^{\mathrm{T}} \geq M$. We can limit ourselves to such policies, and then all states with $\mathbf{x}\mathbf{e}^{\mathrm{T}} \geq M$ are transient. Consequently, for the purpose of finding average optimal policies, we may restrict the state space of $(\mathbf{X}(t), Y(t))$ to the finite set $\mathcal{S} = \{(\mathbf{x}, y) \in \mathbb{N}_0^m \times \Theta | \mathbf{x}\mathbf{e}^{\mathrm{T}} \leq M\}$ for some $M \in \mathbb{N}$. We remark that in the proof of Proposition 4.1 (ii), it is shown how such an M can be found.

4.4.1.2 MDP formulation with bounded transition rates

In this section, we consider the model $\mathfrak{M}(S)$ with $c_e < p\mathbb{E}[L_r]$, and state space $\mathcal{S} = \{(\mathbf{x}, y) \in \mathbb{N}_0^m \times \Theta | \mathbf{x}\mathbf{e}^{\mathrm{T}} \leq M\}$, where M is chosen such that it is optimal to expedite whenever $\mathbf{X}\mathbf{e}^{\mathrm{T}} \geq M$. (By Proposition 4.1, such a finite $M \in \mathbb{N}$ exists.) With a slight abuse of notation, we term the problem of finding an optimal policy for this model as $\mathfrak{M}(S, M)$. In this finite state space, transition rates are bounded and so we can apply the technique of uniformization to transform the problem of finding an optimal expediting policy to discrete time.

Remark 4.1 Without Proposition 4.1, uniformization would not have been possible. Thus, Proposition 4.1, not only facilitates the computation of optimal policies, but is also essential in establishing the structure of optimal policies using an inductive approach based on the dynamic programming recursion. ◇

In each state $(\mathbf{x}, y)$, we take a decision as to whether we expedite the repair of a part if the next event happens to be the arrival of a defective part. We let 1 denote the decision to expedite if a part arrives and let 0 be the decision to not expedite if a part arrives. Thus the action space in state $(\mathbf{x}, y)$ is $\mathcal{A}(\mathbf{x}, y) = \{0, 1\}$ when $\mathbf{x}\mathbf{e}^{\mathrm{T}} < M$ and $\mathcal{A}(\mathbf{x}, y) = \{1\}$ otherwise. Observe that if we take a decision 1 in some state of the system, this does not necessarily imply we will expedite some part, because the next event in the systems may not be the arrival of a defective part.

As uniform transition rate for this MDP, we choose $\Lambda = \lambda_{\max} + M \sum_{i=1}^{m} \mu_i + q_{\max}$. Let $p((\mathbf{x}', y')|(\mathbf{x}, y), a)$ denote the transition probability from state $(\mathbf{x}, y) \in \mathcal{S}$ to $(\mathbf{x}', y') \in \mathcal{S}$ when action $a \in \mathcal{A}(\mathbf{x}, y)$ is taken and note that the time between transitions has an exponential distribution with mean $1/\Lambda$. Without loss of generality, we rescale time

such that $\Lambda = 1$. Then we have:

$$p((\mathbf{x}', y')|(\mathbf{x}, y), a) = \begin{cases} \lambda_y, & \text{if } \mathbf{x}' = \mathbf{x} + \mathbf{e}_1,\ y' = y,\ a = 0; \\ x_m \mu_m, & \text{if } \mathbf{x}' = \mathbf{x} - \mathbf{e}_m,\ y' = y,\ a \in \{0, 1\}; \\ x_i \mu_i, & \text{if } \mathbf{x}' = \mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1},\ y' = y, \\ & a \in \{0, 1\},\ i = 1, ..., m-1; \\ q_{y,y'}, & \text{if } \mathbf{x}' = \mathbf{x},\ y' \neq y,\ a \in \{0, 1\}; \\ \sum_{i=1}^{m}(M - x_i)\mu_i + & \\ \quad q_{\max} - q_y + \lambda_{\max} - \lambda_y, & \text{if } (\mathbf{x}', y') = (\mathbf{x}, y),\ a = 0; \\ \sum_{i=1}^{m}(M - x_i)\mu_i + & \\ \quad q_{\max} - q_y + \lambda_{\max}, & \text{if } (\mathbf{x}', y') = (\mathbf{x}, y) \text{ and } a = 1; \\ 0, & \text{otherwise}, \end{cases} \tag{4.3}$$

where $\mathbf{e}_i$ is the i-th unit vector in dimension m. Regardless of the decision taken, between transitions, an expected penalty cost of $c_p(\mathbf{x}\mathbf{e}^{\mathrm{T}}, y)$ is incurred. Additionally, a cost of c_e is incurred if an arriving defective part is rejected from the system.

Now let $V_n(\mathbf{x}, y)$ denote the optimal total cost function when in state $(\mathbf{x}, y)$ and having n transitions to go and define $V_0(\mathbf{x}, y) \equiv 0$. The finite horizon dynamic programming recursion (Bellman equation) is given by

$$\begin{aligned} V_{n+1}(\mathbf{x}, y) = {} & c_p(\mathbf{x}\mathbf{e}^{\mathrm{T}}, y) + \lambda_y \mathbf{1}_{\{\mathbf{x}\mathbf{e}^{\mathrm{T}} < M\}} \min\{c_e + V_n(\mathbf{x}, y), V_n(\mathbf{x} + \mathbf{e}_1, y)\} \\ & + \lambda_y \mathbf{1}_{\{\mathbf{x}\mathbf{e}^{\mathrm{T}} = M\}}(c_e + V_n(\mathbf{x}, y)) + \textstyle\sum_{i=1}^{m-1} x_i \mu_i V_n(\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}, y) \\ & + x_m \mu_m V_n(\mathbf{x} - \mathbf{e}_m, y) + \textstyle\sum_{i=1}^{m}(M - x_i)\mu_i V_n(\mathbf{x}, y) \\ & + \textstyle\sum_{y' \in \Theta \setminus \{y\}} q_{yy'} V_n(\mathbf{x}, y') + (q_{\max} - q_y + \lambda_{\max} - \lambda_y) V_n(\mathbf{x}, y), \end{aligned} \tag{4.4}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

Remark 4.2 Note that an alternate uniformization constant is given by $\Lambda' = \lambda_{\max} + M\mu_{\max} + q_{\max}$, where $\mu_{\max} = \max_{i \in \{1,\ldots,m\}} \mu_i$. This uniformization constant is smaller, and therefore more suitable if the dynamic programming recursion is used in value iteration algorithms. With this uniformization constant the MDP recursion becomes (we again scale time such that $\Lambda' = 1$):

$$\begin{aligned} V_{n+1}(\mathbf{x}, y) = {} & c_p(\mathbf{x}\mathbf{e}^{\mathrm{T}}, y) + \lambda_y \mathbf{1}_{\{\mathbf{x}\mathbf{e}^{\mathrm{T}} < M\}} \min\{c_e + V_n(\mathbf{x}, y), V_n(\mathbf{x} + \mathbf{e}_1, y)\} \\ & + \lambda_y \mathbf{1}_{\{\mathbf{x}\mathbf{e}^{\mathrm{T}} = M\}}(c_e + V_n(\mathbf{x}, y)) \\ & + \textstyle\sum_{i=1}^{m-1} x_i \mu_i V_n(\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}, y) + x_m \mu_m V_n(\mathbf{x} - \mathbf{e}_m, y) \\ & + \textstyle\sum_{i=1}^{m} x_i(\mu_{\max} - \mu_i) V_n(\mathbf{x}, y) + (M - \sum_{i=1}^{m} x_i)\, \mu_{\max} V_n(\mathbf{x}, y) \\ & + \textstyle\sum_{y' \in \Theta \setminus \{y\}} q_{yy'} V_n(\mathbf{x}, y') + (q_{\max} - q_y + \lambda_{\max} - \lambda_y) V_n(\mathbf{x}, y). \end{aligned} \tag{4.5}$$

In this section, we will work with the formulation with the computationally 'less efficient' Λ so that we can reuse some results in the literature to prove structural

properties of optimal policies. In the numerical section, §4.6, we use the formulation in because a smaller uniformization constant leads to quicker convergence of value iteration algorithms (e.g Kulkarni, 1999). ◇

To analyze the value function $V_n(\mathbf{x}, y)$ in §4.4.1.3, we employ the event based dynamic programming approach introduced by Koole (1998, 2006). To this end, let $\mathcal{V}$ denote the set of all functions $v : \mathcal{S} \to \mathbb{R}$ and let $f, f_1, ..., f_{m+2} \in \mathcal{V}$. We define the following operators $\mathbb{T}_{\text{cost}}, \mathbb{T}_{\text{AC}(i)}, \mathbb{T}_{\text{TD}(i)}, \mathbb{T}_{\text{D}(i)}, \mathbb{T}_{\text{env}} : \mathcal{V} \to \mathcal{V}$, $\mathbb{T}_{\text{unif}} : \mathcal{V}^{m+2} \to \mathcal{V}$.

$$\mathbb{T}_{\text{cost}} f(\mathbf{x}, y) = c_p(\mathbf{x}\mathbf{e}^{\text{T}}, y) + f(\mathbf{x}, y) \tag{4.6}$$

$$\begin{aligned}\mathbb{T}_{\text{AC}(i)} f(\mathbf{x}, y) = {} & \mathbf{1}_{\{\mathbf{x}\mathbf{e}^{\text{T}} < M\}} \min\{c_e + f(\mathbf{x}, y), f(\mathbf{x} + \mathbf{e}_i, y)\} \\ & + \mathbf{1}_{\{\mathbf{x}\mathbf{e}^{\text{T}} = M\}}(c_e + f(\mathbf{x}, y))\end{aligned} \tag{4.7}$$

$$\mathbb{T}_{\text{TD}(i)} f(\mathbf{x}, y) = \frac{x_i}{M} f(\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}, y) + \frac{M - x_i}{M} f(\mathbf{x}, y) \tag{4.8}$$

$$\mathbb{T}_{\text{D}(i)} f(\mathbf{x}, y) = \frac{x_i}{M} f(\mathbf{x} - \mathbf{e}_i, y) + \frac{M - x_i}{M} f(\mathbf{x}, y) \tag{4.9}$$

$$\mathbb{T}_{\text{env}} f(\mathbf{x}, y) = \sum_{y \in \Theta \setminus \{y\}} q_{yy'} f(\mathbf{x}, y') + (q_{\max} - q_y + \lambda_{\max} - \lambda_y) f(\mathbf{x}, y) \tag{4.10}$$

$$\mathbb{T}_{\text{unif}}(f_1, \cdots, f_{m+2})(\mathbf{x}, y) = \lambda_y f_1(\mathbf{x}, y) + \sum_{i=1}^{m} M \mu_i f_{i+1}(\mathbf{x}, y) + f_{m+2}(\mathbf{x}, y) \tag{4.11}$$

These operators are variations to operators defined by Koole (1998, 2004, 2006) and are originally intended to model various common queueing mechanisms such as arrival control ($\mathbb{T}_{\text{AC}(i)}$), transfer departures from multi-server tandem queues ($\mathbb{T}_{\text{TD}(i)}$), and departures from multi-server queues ($\mathbb{T}_{\text{D}(i)}$), while the operators $\mathbb{T}_{\text{cost}} f(\mathbf{x}, y)$, $\mathbb{T}_{\text{env}}$ and $\mathbb{T}_{\text{unif}}$ are mainly convenient for bookkeeping. The Bellman recursion for our MDP, (4.4), can now be written succinctly as

$$\begin{aligned}V_{n+1}(\mathbf{x}, y) = \mathbb{T}_{\text{cost}} \mathbb{T}_{\text{unif}} \big[& \mathbb{T}_{\text{AC}(1)} V_n(\mathbf{x}, y), \mathbb{T}_{\text{TD}(1)} V_n(\mathbf{x}, y), \cdots, \mathbb{T}_{\text{TD}(m-1)} V_n(\mathbf{x}, y), \\ & \mathbb{T}_{\text{D}(m)} V_n(\mathbf{x}, y), \mathbb{T}_{\text{env}} V_n(\mathbf{x}, y) \big].\end{aligned} \tag{4.12}$$

This formulation of the MDP recursion is convenient because the propagation of value function properties over n can be analyzed through the propagation properties of operators, for which results are available in literature.

We remark that the operators used to rewrite the MDP recursion reveal that the MDP we are dealing with is equivalent to an admission control problem for a tandem line of ample exponential server queues. A similar equivalence is exploited by Song and Zipkin (2009) in finding effective means to evaluate heuristic policies.

4.4.1.3 Average optimal expediting policies

To characterize average optimal policies, we study properties of the value function and how these properties propagate through recursion (4.12). We define the first order difference operator with respect to x_i, Δ_i, as $\Delta_i f(\mathbf{x}, y) = f(\mathbf{x} + \mathbf{e}_i, y) - f(\mathbf{x}, y)$. We distinguish the following subsets of $\mathcal{V}$:

$$\mathcal{I}(i) = \{f \in \mathcal{V} | f(\mathbf{x}, y) \leq f(\mathbf{x} + \mathbf{e}_i, y)\} \tag{4.13}$$

$$\mathcal{C}(i) = \{f \in \mathcal{V} | \Delta_i f(\mathbf{x}, y) \leq \Delta_i f(\mathbf{x} + \mathbf{e}_i, y)\} \tag{4.14}$$

$$\mathcal{UI} = \{f \in \mathcal{V} | f(\mathbf{x} + \mathbf{e}_{i+1}, y) \leq f(\mathbf{x} + \mathbf{e}_i, y), i = 1, ..., m-1\} \tag{4.15}$$

$$\mathcal{SM}(i,j) = \{f \in \mathcal{V} | \Delta_i f(\mathbf{x}, y) \leq \Delta_i f(\mathbf{x} + \mathbf{e}_j, y)\}. \tag{4.16}$$

In (4.13)-(4.16), it is understood that the inequalities that characterize each set must hold when the arguments on both sides of the inequality exist in $\mathcal{S}$. $\mathcal{I}(i)$ and $\mathcal{C}(i)$ are the sets of non-decreasing and convex functions with respect to x_i respectively. $\mathcal{UI}$ is the set of upstream increasing functions as introduced in Koole (2004) and renamed in Koole (2006). The set $\mathcal{SM}(i,j)$ consists of functions with a specific supermodularity property. Finally, define $\mathcal{F}$ as

$$\mathcal{F} = \left(\bigcap_{i=1}^{m} \mathcal{I}(i) \right) \cap \left(\bigcap_{j=2}^{m} \mathcal{SM}(1,j) \right) \cap \mathcal{UI} \cap \mathcal{C}(1). \tag{4.17}$$

Lemma 4.2 *The following statements hold:*

(i) The function $g \in \mathcal{V}$ defined by $g(\mathbf{x}, y) = c_p(\mathbf{x}\mathbf{e}^T, y)$ is a member of $\mathcal{F}$, i.e., $g \in \mathcal{F}$.

(ii) If $f \in \mathcal{F}$ then $\mathbb{T}_{\text{cost}} f(\mathbf{x}, y), \mathbb{T}_{\text{AC}(1)} f(\mathbf{x}, y), \mathbb{T}_{\text{D}(m)} f(\mathbf{x}, y), \mathbb{T}_{\text{env}} f(\mathbf{x}, y) \in \mathcal{F}$ and $\mathbb{T}_{\text{TD}(i)} f(\mathbf{x}, y) \in \mathcal{F}$ for $i = 1, ..., m-1$.

(iii) If $f_j \in \mathcal{F}$ for $j = 1, ..., m+2$, then $\mathbb{T}_{\text{unif}}(f_1, ..., f_{m+2})(\mathbf{x}, y) \in \mathcal{F}$.

The proof of this lemma is in Appendix 4.B.3. The properties of functions in $\mathcal{V}$ that are shown to propagate through operators (4.6)-(4.11) in Lemma 4.2, imply structure on the optimal policy. To state the next lemma, we introduce some notation. Let $\mathbf{x}^{(-1)}$ denote the vector $\mathbf{x}$ with its first component set to 0, i.e., $\mathbf{x}^{(-1)} = (0, x_2, \ldots, x_m)$. The next lemma explains how the optimal policy at transition epoch n is related to properties of V_{n-1}.

Lemma 4.3 *If $V_{n-1} \in \mathcal{F}$, then, at transition epoch n, there are state dependent thresholds $T_n\left(\mathbf{x}^{(-1)}, y\right)$ such that it is optimal to expedite the repair of an arriving*

part at transition epoch n if $X_1(t_n) \geq T_n\left(\mathbf{X}^{(-1)}(t_n), Y(t_n)\right)$, where t_n is the time corresponding to transition epoch n. Furthermore the thresholds $T_n\left(\mathbf{x}^{(-1)}, y\right)$ satisfy the following monotonicity property: $\Delta_i T_n\left(\mathbf{x}^{(-1)}, y\right) \leq 0$, for $i = 2, ..., m$.

PROOF: Let $V_{n-1} \in \mathcal{F}$. The fact that a state dependent threshold policy is optimal at decision epoch n follows immediately from the fact that $V_{n-1} \in \mathcal{C}(1)$ and Koole (2006), Theorem 8.1. Because $V_{n-1} \in \bigcap_{i=2}^{m} \mathcal{SM}(1, i)$, this threshold is non-increasing in x_2 to x_m, again by Koole (2006), Theorem 8.1. This can be written as $T_n(\mathbf{x} + \mathbf{e}_i, y) \leq T_n(\mathbf{x}, y)$ for $i = 2, ..., m$, and by subtracting $T_n(\mathbf{x}, y)$ from both sides we obtain $\Delta_i T_n(\mathbf{x}, y) \leq 0$. □

An alternative interpretation of Lemma 4.3 is that the optimal policy at transition epoch n (under the stated condition) is a switching curve between expediting and not expediting repair. This switching curve is decreasing in x_i for $i = 2, ..., m$. Figure 4.2 illustrates two such switching curves. In part (a) of the figure, for a given x_2, it is optimal to expedite repair if x_1 is on or above the shown line. In part (b) of the figure, for given (x_2, x_3) it is optimal to expedite repair if x_1 is on or above the shown surface.

The policy described in Lemma 4.3 can also be reinterpreted as a state dependent expedite-up-to policy. To see this, define $IP_e(t) = S - \mathbf{X}(t)\mathbf{e}^{\mathrm{T}}$, and note that this can be interpreted as the expedited inventory position: on-hand inventory minus backorders plus outstanding orders arriving within the expedited lead time ℓ_e. The optimal policy is now to expedite parts to retain $IP_e(t_n)$ at or above the level $S - T_n\left(\mathbf{X}^{(-1)}(t_n), Y(t_n)\right)$. Thus the resulting policy is a state dependent version of the dual-index policy (Veeraraghavan and Scheller-Wolf, 2008; Arts et al., 2011, consider state independent dual-index policies), where regular and emergency inventory positions are both kept at or above their order-up-to levels. Note however, that the regular order-up-to level was assumed to be S from the start as we are dealing with a closed loop system. Without this fixed base-stock level, a state dependent dual-index replenishment policy need not be optimal.

The main result of this section is that average optimal policies also have the structure described in Lemma 4.3.

Theorem 4.1 *Consider the model $\mathfrak{M}(S)$. If $c_e \geq p\mathbb{E}[L_r]$, then it is average optimal to never expedite repair. If $c_e < p\mathbb{E}[L_r]$, then there are state dependent threshold levels $T\left(\mathbf{x}^{(-1)}, y\right) \in \mathbb{N}_0$ such that it is average optimal to expedite the repair of an arriving defective part at time t if $X_1(t) \geq T\left(\mathbf{X}^{(-1)}(t), Y(t)\right)$. Furthermore these threshold levels $T\left(\mathbf{x}^{(-1)}, y\right)$ satisfy the property in Lemma 4.3, i.e., $\Delta_i T\left(\mathbf{x}^{(-1)}, y\right) \leq 0$ for $i = 2, ..., m$.*

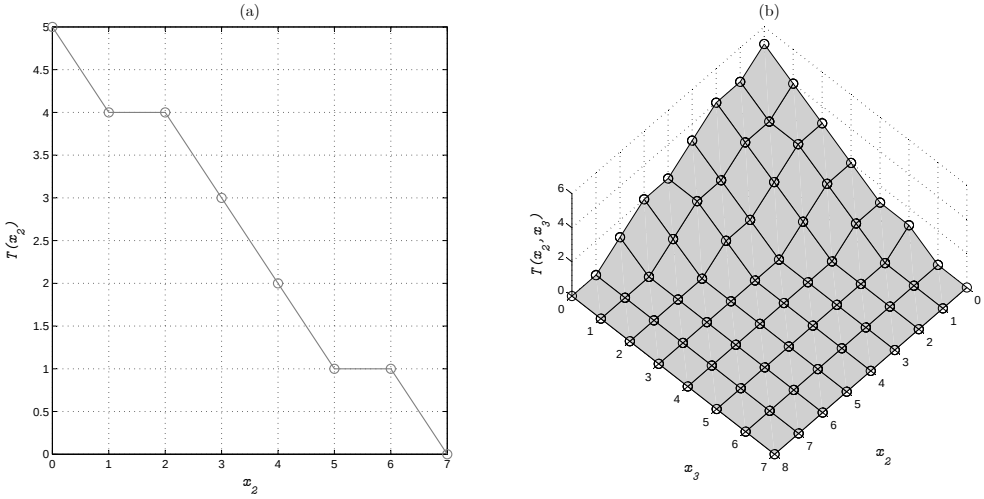

Figure 4.2 Part (a) shows the state dependent threshold for $n = 593$ in the case where L_r has an Erlang(2) distribution ($m = 2$). Part (b) shows the state dependent thresholds for $n = 1513$ when L_r is Erlang(3) distributed ($m = 3$). Both cases are based on the problem instance with $|\Theta| = 1$, $\lambda_1 = 1$, $\mathbb{E}[L_r] = 4$, $\ell_e = 2$, $c_e = 8$, $p = 10$, and $S = 8$. In both cases, n coincides with the iteration in which average optimal policies are found within some precision.

PROOF: The first part of the theorem is simply a restatement of part (ii) of Proposition 4.1. If $c_e < p\mathbb{E}[L_r]$, we may apply Proposition 4.1.(ii) to truncate the state space at a finite M and optimal solutions to $\mathfrak{M}(S, M)$ will coincide to optimal solutions to $\mathfrak{M}(S)$ provided M is sufficiently large. Therefore, let us consider $\mathfrak{M}(S, M)$. Observe that state $(\mathbf{0}, y)$ for any $y \in \Theta$ is reachable from any other state for any policy, so this MDP is unichain. Furthermore, since under any policy, there are transitions from state $(\mathbf{0}, y)$ to itself, this MDP is aperiodic. By Theorem 8.5.4 of Puterman (1994), $\max_{(\mathbf{x},y)\in\mathcal{S}}(V_{n+1}(\mathbf{x}, y) - V_n(\mathbf{x}, y)) - \min_{(\mathbf{x},y)\in\mathcal{S}}(V_{n+1}(\mathbf{x}, y) - V_n(\mathbf{x}, y))$ converges to the optimal average costs as $n \to \infty$. Now for each n, a policy of the form described in Lemma 4.3 is optimal because $V_0 \in \mathcal{F}$ and so, by induction using Lemma 4.2, so are V_n for $n \in \mathbb{N}$. Finally since both the state and action space of this MDP are finite, there are finitely many policies that satisfy Lemma 4.3, and at least one of them will be found infinitely often throughout recursion (4.12). Such a repair expediting policy is average optimal. □

Theorem 4.1 also answers a question and conjecture posed by Song and Zipkin (2009, p. 371): "Are there any systems for which some policy of the form above is in fact

optimal?". The policy Song and Zipkin (2009) propose is exactly the policy described in Theorem 4.1 for the special case that $m = 1$. For $m \geq 2$ one obtains a generalized form of this policy.

4.4.1.4 Infinite horizon discounted optimal expediting policies

The same policy structure results hold for the case where we are interested in the infinite horizon discounted cost criterion. Let $\beta > 0$ be the discount rate. Proposition 4.1 continues to hold with $p\mathbb{E}[L_r]$ replaced by the expected discounted penalty costs over an interval of length L_r:

$$\mathbb{E}_{L_r}\left[\int_0^{L_r} pe^{-\beta t}dt\right] = \frac{p}{\beta} - \frac{p}{\beta}\mathbb{E}\left[e^{-\beta L_r}\right].$$

This holds in general for all non-negative distributions that might model L_r. In our particular model, the Laplace-Stieltjes transform of L_r is given by:

$$\mathbb{E}\left[e^{-\beta L_r}\right] = \prod_{i=1}^{m} \frac{\mu_i}{\mu_i + \beta}.$$

The MDP recursion can be written in exactly the same manner as before except that $\mathbb{T}_{\text{cost}}$, needs to be changed to $\mathbb{T}^{\beta}_{\text{cost}} : \mathcal{V} \to \mathcal{V}$ with

$$\mathbb{T}^{\beta}_{\text{cost}} f(\mathbf{x}, y) = \frac{c_p(\mathbf{x}\mathbf{e}^{\mathrm{T}}, y)}{\Lambda + \beta} + \frac{\Lambda}{\Lambda + \beta} f(\mathbf{x}, y).$$

It is readily verified that $\mathbb{T}^{\beta}_{\text{cost}}$ propagates the same properties as $\mathbb{T}_{\text{cost}}$, that is, if $f \in \mathcal{F}$ then also $\mathbb{T}^{\beta}_{\text{cost}} f(\mathbf{x}, y) \in \mathcal{F}$. With this change, it is easy to verify that Theorem 4.1 still holds, again with $p\mathbb{E}[L_r]$ changed to $\frac{p}{\beta} - \frac{p}{\beta}\mathbb{E}\left[e^{-\beta L_r}\right]$.

Theorem 4.2 *Consider the infinite horizon discounted cost criterion for model $\mathfrak{M}(S)$ with discount rate β. If $c_e \geq \frac{p}{\beta} - \frac{p}{\beta}\mathbb{E}\left[e^{-\beta L_r}\right] = \frac{p}{\beta} - \frac{p}{\beta}\prod_{i=1}^{m}\frac{\mu_i}{\mu_i+\beta}$, then it is β-discounted optimal to never expedite repair. If $c_e < \frac{p}{\beta} - \frac{p}{\beta}\mathbb{E}\left[e^{-\beta L_r}\right] = \frac{p}{\beta} - \frac{p}{\beta}\prod_{i=1}^{m}\frac{\mu_i}{\mu_i+\beta}$, then there are state dependent threshold levels $T\left(\mathbf{x}^{(-1)}, y\right) \in \mathbb{N}_0$ such that it is β-discounted optimal to expedite repair at time t if $X_1(t) \geq T\left(\mathbf{X}^{(-1)}(t), Y(t)\right)$. Furthermore these threshold levels $T\left(\mathbf{x}^{(-1)}, y\right)$ satisfy the property in Lemma 4.3, i.e., $\Delta_i T\left(\mathbf{x}^{(-1)}, y\right) \leq 0$ for $i = 2, ..., m$.*

Remark 4.3 Numerical results indicate that optimal expediting thresholds satisfy additional monotonicity properties, namely that

$$-1 \leq \Delta_2 T\left(\mathbf{x}^{(-1)}, y\right) \leq \Delta_3 T\left(\mathbf{x}^{(-1)}, y\right) \leq \ldots \leq \Delta_m T\left(\mathbf{x}^{(-1)}, y\right) \leq 0, \qquad (4.18)$$

implying in particular that $T\left(\mathbf{x}^{(-1)}+\mathbf{e}_i, y\right) \leq T\left(\mathbf{x}^{(-1)}+\mathbf{e}_{i+1}, y\right)$ for $i \in \{2,\ldots,m\}$. This property essentially formalizes that the optimal expediting policy is more sensitive to late to arrive repair orders, than it is to soon to arrive repair orders. To prove that this property holds, it suffices to show that

$$\Delta_1 V_n(\mathbf{x}+\mathbf{e}_i, y) \geq \Delta_1 V_n(\mathbf{x}+\mathbf{e}_{i+1}, y) \tag{4.19}$$

for all $(\mathbf{x}, y)$ in $\mathcal{S}$, $n \in \mathbb{N}$ and $i \in \{1,\ldots,m-1\}$, but we have only been able to establish it for $i \in \{2,\ldots,m-1\}$. The property in (4.19) might be called upstream convexity. Unfortunately, we have been unable to prove this property. The difficulty lies in proving that this property is propagated by $\mathbb{T}_{\mathrm{TD}(1)}$. Ghoneim and Stidham Jr (1985) and Moreta and Ziedins (1998) have also reported that they were unable to prove this property for very similar systems. At the same time, they too were unable to find a counterexample. ◇

4.4.2 Turn-around stock optimization

In the analysis in the previous sections, we have considered problem $\mathfrak{M}(S)$, i.e., we have considered S to be a fixed constant that was determined at $t = 0$, and have focussed on using the expedition decision to minimize expedition and backorder penalty costs. Now, we focus on the joint optimization of the turn-around stock S and the expediting policy. For brevity of exposition, we only discuss the average cost criterion in this section.

To facilitate presentation, we first present some notation: We let $C(S)$ denote the optimal average expediting and backorder penalty costs per time unit for a turn-around stock of size S, i.e., $C(S) = \lim_{n\to\infty} V_n(\mathbf{0}, 1)/n$ is the optimal cost associated with $\mathfrak{M}(S)$. Furthermore, we let $C_{\mathrm{tot}}(S) := hS + C(S)$ denote the total cost rate associated with a turn-around stock of S if an optimal repair expediting policy is used.

Whenever we drop the time index of a stochastic process, we are referring to the process in steady state, e.g. $\mathbb{P}\{Y = y\} = \lim_{t\to\infty} \mathbb{P}\{Y(t) = y\}$. We let the random variable $D(L)$ denote demand in an interval of length $L \geq 0$ when the modulating chain of demand is in steady state, i.e.,

$$\mathbb{P}\{D(L) \leq k\} = \sum_{y\in\Theta} \mathbb{P}\{Y = y\}\mathbb{P}\left\{D^y_{t,t+L} \leq k\right\}.$$

A lower bound of $C_{\mathrm{tot}}(S)$ is given by the average holding and backorder penalty cost rates of the system with turn-around stock S under the feasible policy of expediting everything against zero expediting cost:

$$C_{LB}(S) := hS + p\mathbb{E}\left[(D(\ell_e) - S)^+\right].$$

When we do include the expediting costs, we obtain an upper bound for $C_{\text{tot}}(S)$:

$$C_{UB}(S) := C_{LB}(S) + c_e\bar{\lambda}.$$

Here, $\bar{\lambda} = \sum_{y\in\Theta} \lambda_y \mathbb{P}\{Y = y\}$ is the long run average demand per time period. Let $S^* := \text{argmin}_{S\in\mathbb{N}_0} C_{\text{tot}}(S)$ denote the optimal turn-around stock. An upper bound to $C_{\text{tot}}(S^*)$ is obtained by minimizing $C_{UB}(S)$. The S that minimizes $C_{UB}(S)$ (as well as $C_{LB}(S)$) can be easily found as $C_{UB}(S)$ is convex. We denote this minimizer $\hat{S}$ and it is the smallest integer that satisfies the newsvendor inequality

$$\mathbb{P}\left\{D(\ell_e) \le \hat{S}\right\} \ge \frac{p-h}{p}. \tag{4.20}$$

Since $C_{LB}(S)$ is convex, it is easy to find the greatest $S \le \hat{S}$ and smallest $S \ge \hat{S}$ such that $C_{LB}(S) \ge C_{UB}(\hat{S})$. This will provide lower and upper bounds respectively on S^*.

Proposition 4.2 *The optimal turn-around stock, S^*, that minimizes $C_{tot}(S)$ is bounded as $S_{LB} \le S^* < S_{UB}$, where S_{LB} and S_{UB} are given by*

$$S_{LB} = \max\left\{\{0\} \cup \min\{x \in \mathbb{N}_0, x \le \hat{S} : C_{LB}(x) \ge C_{UB}(\hat{S})\}\right\} \tag{4.21}$$

$$S_{UB} = \min\{x \in \mathbb{N}_0, x \ge \hat{S} : C_{LB}(x) \ge C_{UB}(\hat{S})\}, \tag{4.22}$$

Furthermore, if $C(S) \le h$ for some $S \in \mathbb{N}_0$, then $S^* \le S$.

PROOF: The bounds established by S_{LB} and S_{UB} follow directly from the analysis preceding Proposition 4.2. To verify the last statement, observe that $C(S)$ is non-negative and decreasing and that $\Delta C_{\text{tot}}(S) = h + \Delta C(S)$. Combining these facts implies that if $C(S) \le h$, then $\Delta C_{\text{tot}}(S) \ge 0$ and so $S^* < S$. □

Proposition 4.2 gives us bounds that can be used to minimize $C_{\text{tot}}(S)$ by enumeration.

A natural question is whether $C(S)$ is convex in S as this would make optimization easy. Unfortunately $C(S)$ is not convex in S as can be verified by considering the problem instance with Poisson demand with rate $\lambda = 10$, $\ell_e = 1$, $\mathbb{E}[L_r] = 3$, $m = 1$, $p = 10$ and $c_e = 15$. In this case, $C(S)$ can be obtained exactly without dynamic programming using the results in Moinzadeh and Schmidt (1991) and Song and Zipkin (2009). (In §4.5.1.1, we also show how to compute $C(S)$ without dynamic programming for this instance.) For this instance it can be verified that $C(20) - 2C(19) + C(18) \le -0.03$, showing that $C(S)$ is not convex in general; see also Figure 4.3. The non-convexity of $C(S)$ does affect the unimodality of $C_{\text{tot}}(S)$ but only in rather extreme cases. Figure 4.4 presents such a case. In §4.6 we present

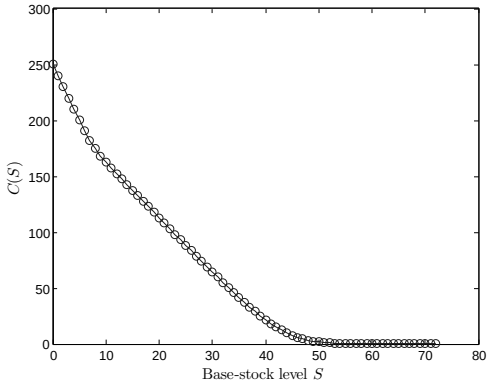

Figure 4.3 Consider the problem instance with Poisson demand with rate $\lambda = 10$, $\ell_e = 1$, $m = 1$, $\mathbb{E}[L_r] = 3$, $p = 10$ and $c_e = 15$. This figure shows the optimal cost for expediting and backordering as a function of the turn-around stock $C(S)$. Note in particular the non-convexity around $S = 19$.

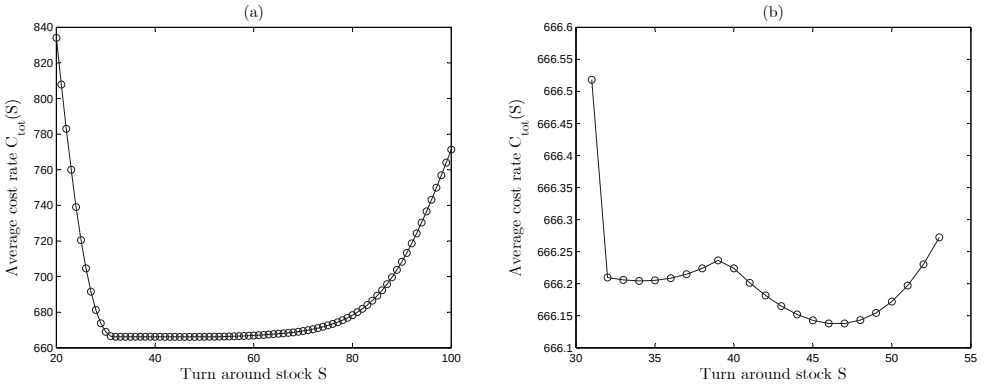

Figure 4.4 Consider the problem instance with Poisson demand with rate $\lambda = 3.43$, $\ell_e = 8$, $m = 1$, $\mathbb{E}[L_r] = 15$, $p = 37.2$ and $c_e = 116$. This figure shows $C_{\text{tot}}(S)$ for this instance. Although the fact that $C_{\text{tot}}(S)$ is not unimodal is not immediately apparent from sub-figure (a), it is apparent from sub-figure (b). Any small deviation of any of the problem parameters will make $C_{\text{tot}}(S)$ unimodal.

numerical work for instances as they are typically encountered in practice. For all these instances, $C(S)$ is convex and $C_{\text{tot}}(S)$ is unimodal. In fact, a cursory look at Figures 4.3 and 4.4.a does not immediately reveal that $C(S)$ and $C_{\text{tot}}(S)$ are not convex. This is typical for all counterexamples we have found.

4.4.2.1 Trading off safety stock and safety time

In this subsection, we formally show that as the turn-around stock increases, the need for expediting decreases and vice versa. To formalize this, we need some additional notation. Let $\mathcal{W}$ be set of all functions $w : \mathcal{S} \times \mathbb{N}_0 \to \mathbb{R}$. The extra argument corresponds to the turn-around stock. We make the dependence of $V_n(\mathbf{x}, y)$ on S explicit by writing $V_n(\mathbf{x}, y, S)$. Note that operators (4.6)-(4.11) are also mappings from $\mathcal{W} \to \mathcal{W}$. We start with a sub-modularity of $V_n(\mathbf{x}, y, S)$ with respect to x_1 and S.

Lemma 4.4 *For all $n \in \mathbb{N}_0$, $(x, y) \in \mathcal{S}$ and $S < S_{UB}$,*

$$\Delta_1 V_n(\mathbf{x}, y, S) \geq \Delta_1 V_n(\mathbf{x}, y, S + 1) \tag{4.23}$$

where $\mathcal{S}$ is chosen with an appropriate M that is common for all $S < S_{LB}$

The proof of this lemma is in the appendix and follows an inductive approach. Now we can formally state that as the turn-around stock increases, the need for expediting decreases and vice versa.

Proposition 4.3 *Let $T_S\left(\mathbf{x}^{(-1)}, y\right)$ denote the expediting threshold that is average optimal under a turn-around stock level of S at $(\mathbf{x}, y) \in \mathcal{S}$. Then $T_S\left(\mathbf{x}^{(-1)}, y\right) \leq T_{S+1}\left(\mathbf{x}^{(-1)}, y\right)$ for all $(\mathbf{x}, y) \in \mathcal{S}$ and $S \in \mathbb{N}_0$, that is, when the turn-around stock increases, the need for expediting decreases.*

PROOF: Because the expediting decision is taken whenever $\min\{V_n(\mathbf{x}, y, S) + c_e, V_n(\mathbf{x} + \mathbf{e}1, y, S)\} = V_n(\mathbf{x}, y) + c_e$, which is equivalent to $\Delta_1 V_n(\mathbf{x}, y, S) \geq c_e$, it is clear that Lemma 4.4 implies the result. □

The interpretation of Proposition 4.3 is that the possibility to expedite acts as kind of safety time, that can be used in stead of (safety) stock to lower the risk of backorders.

4.5. E-WDT Heuristic

In the previous section, we have analyzed an exact solution to our problem. However, finding the optimal solution involves solving an MDP which suffers from the curse of dimensionality for each S between S_{LB} and S_{UB}. Furthermore, the optimal expediting policy is rather intricate, depending on the entire vector of repair that will not arrive within the expedited lead time. In this section, we describe a heuristic for our model that involves an expediting policy that is much easier to interpret and that does not

impose the same computational burden. We call this heuristic the E-WDT heuristic for reasons that will become clear later. This section is organized in the same fashion as the previous section: First we discuss heuristic expediting policies in §4.5.1 and then we discuss the heuristic optimization of the turn-around stock in §4.5.2.

4.5.1 World driven threshold policies

Computing the state dependent optimal threshold levels quickly becomes computationally prohibitive as m increases. A plausible heuristic policy is to aggregate all orders in $\mathbf{X}(t)$ and to put a threshold expediting level, $T(y)$, on their sum $\mathbf{X}(t)\mathbf{e}^{\mathrm{T}}$. This threshold will then only depend on $Y(t)$ and so, borrowing the terminology of Zipkin (2000), we call such a policy a world driven threshold (WDT) policy. It is readily verified that the WDT policy satisfies the monotonicity property in Theorem 4.1 that $\Delta_i T\left(\mathbf{x}^{(-1)}, y\right) \leq 0$. Indeed, observe that the thresholds $(T^{WDT}(\mathbf{x}, y))$ of a WDT policy satisfy:

$$\Delta_i T^{WDT}(\mathbf{x}, y) = \begin{cases} -1, & \text{if } T(\mathbf{x}, y) > 0; \\ 0, & \text{otherwise.} \end{cases}$$

This is shown graphically in Figure 4.5, where the optimal thresholds are shown with the best WDT thresholds. As before, the most convenient way to interpret Figure 4.5 is to think of it as a switching curve: If x_1 is on or above the shown line for some x_2, then expedite the repair, otherwise do not expedite repair.

For $m > 1$, finding the best WDT policy is about as difficult as finding an optimal policy since the stationary distribution of $\mathbf{X}(t)$ under such a policy still requires the evaluation of an $m + 1$ dimensional Markov chain. A notable exception, that we discuss in §4.5.1.1, occurs when demand is a stationary Poisson process, i.e., $|\Theta| = 1$.

In general, for $|\Theta| > 1$ and $c_e < p\mathbb{E}[L_r]$, we propose the following heuristic way of finding a good WDT policy. In stead of working with the $(m+1)$-dimensional space, move to two-dimensional space by approximating L_r by a single exponential phase with the same mean $\mu_1 = \mu = 1/E[L_r]$. Then we are left with a two-dimensional space for which we can easily solve the resulting MDP to optimality using any common algorithm to solve finite state and action space MDPs such as value iteration, policy iteration or linear programming.

The WDT policies that result from this procedure are not necessarily optimal within the class of WDT policies and the computed cost is not exact but an approximation. Since the system under study is equivalent to a type of ample server queue, we may expect this approximation to be quite accurate. In the next subsection, we show that this approach is exact for Poisson demand and in §4.6, we provide numerical evidence that WDT policies that are found in this manner perform exceptionally well compared

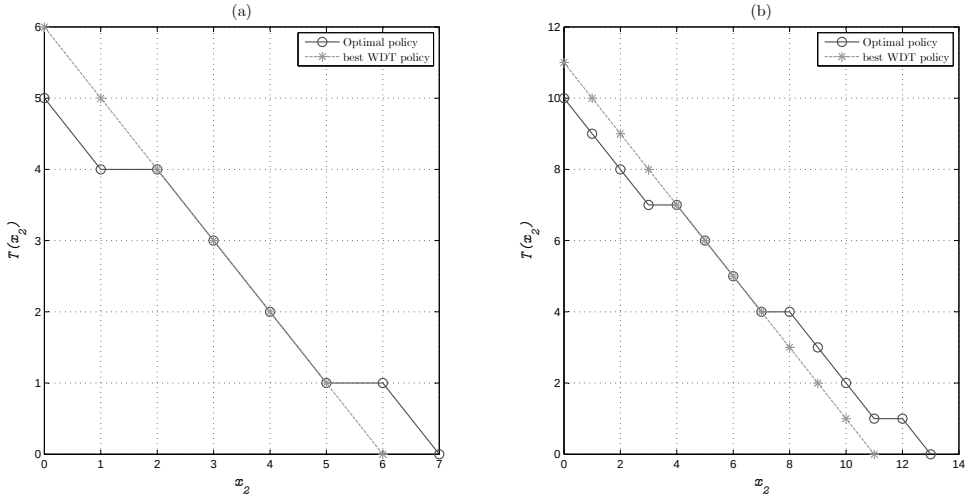

Figure 4.5 Optimal and heuristic policies. Part (a) shows the optimal state dependent threshold in conjunction with the best WDT policy for the case with $\lambda = 1$, $\mathbb{E}[L_r] = 4$, $\ell_e = 2$, $p = 10$, $c_e = 8$, $S = 8$ and $m = 2$. Part (b) shows the optimal state dependent threshold for in conjunction with the best heuristic policy for the same case except $S = 12$.

to optimal policies under Markov modulated Poisson demand. Furthermore, the cost approximations of this method are also very accurate.

4.5.1.1 Special case: Poisson demand

Now we briefly consider the evaluation of WDT policies for the special case where $|\Theta| = 1$, and we are dealing with stationary Poisson demand. In this case, the evaluation of a WDT policy can be done exactly for any distribution of L_r in closed form using the results of Song and Zipkin (2009). (In this context, it might be appropriate to refer to a WDT policy, simply as a threshold policy. For convenience, we use the name WDT policy also in this context.) Alternatively, one may simply observe that $\mathbf{X}(t)\mathbf{e}^{\mathrm{T}}$, under such a policy, has the same stationary distribution as the number of customers in a $M/G/c/c$ queue, where the number of servers c is set equal to the threshold T and the service time is distributed as L_r. In this equivalence, a customer being blocked from the queue because all T servers are busy corresponds to a repair being expedited because there are T or more parts that will not arrive within ℓ_e. The average expediting and backorder penalty cost rate for such a policy

with threshold level T and base-stock level S, $C(T|S)$ is therefore given by:

$$C(T|S) = \lambda c_e B(T, \lambda\mathbb{E}[L_r]) + \sum_{x=0}^{T} c_p(x|S) \frac{(\lambda\mathbb{E}[L_r])^x/x!}{\sum_{k=0}^{T}(\lambda\mathbb{E}[L_r])^k/k!} \tag{4.24}$$

where $B(c,\rho) = \frac{\rho^c/c!}{\sum_{k=0}^{c}\rho^k/k!}$ is the Erlang loss function with c servers and traffic intensity ρ, λ is the intensity of the Poisson demand process, and $c_p(x|S) = c_p(x,1|S)$. Expression (4.24) also reveals that the performance of a WDT policy is insensitive to the distribution of L_r for the special case of Poisson demand. This insensitivity does not hold for Markov modulated Poisson demand. In the numerical section however, we provide evidence that the performance evaluation of a WDT policy is nearly insensitive to the exact distribution of L_r for Markov modulated Poisson demand processes.

4.5.2 Heuristic optimization of the turn-around stock: The E-WDT heuristic

In §4.5.1, we discussed a heuristic to obtain a good WDT policy for a given turn-around stock S, namely by finding the optimal expediting policy after replacing L_r by a single exponential phase with the same mean. (This is of course exact if L_r happens to be exponentially distributed.) The heuristic for joint optimization is based on this idea.

Let $C_{\mathrm{E}}(S)$ denote the optimal expediting and penalty cost rate when L_r has an exponential distribution with mean μ^{-1} and the turn-around stock is S. (The E stands for exponential distribution, as L_r has an exponential distribution in C_{E}.) The heuristic we propose is to minimize $C_{\text{E-tot}}(S) = hS + C_{\mathrm{E}}(S)$ with $\mu^{-1} = \mathbb{E}[L_r]$ using a greedy algorithm such as golden section search. We call this heuristic the E-WDT heuristic. (E stands for exponential and WDT stands for world driven threshold.) There is no guarantee that a greedy search of $C_{\text{E-tot}}(S)$ yields the global optimum of $C_{\text{E-tot}}(S)$ because $C_{\mathrm{E}}(S)$ need not be convex as shown in §4.4.2. However, we have been unable to construct an example $C_{\text{tot}}(S)$ that is not unimodal and this serves as an indication that a greedy algorithm can work well. We let S_{E} denote the base-stock level that is found by applying the E-WDT heuristic and $T_{\mathrm{E}}(y)$ denote the corresponding expediting thresholds that depend only on $y \in \Theta$.

We emphasize that in general $C_{\mathrm{E}}(S_{\mathrm{E}})$ does not represent the real backordering and expediting cost of applying the WDT policy with $T_{\mathrm{E}}(y)$, because these costs are not insensitive to the distribution of L_r unless demand is a Poisson process. Let $C_{\mathrm{R}}(S)$ denote the real backordering and expediting costs when applying the WDT policy found by assuming L_r is exponential to the real system where L_r is not necessarily exponential. Similarly, let $C_{\text{R-tot}}(S) = hS + C_{\mathrm{R}}(S)$. We provide numerical evidence in §4.6.3 that $C_{\mathrm{E}}(S_{\mathrm{E}}) \approx C_{\mathrm{R}}(S_{\mathrm{E}})$.

4.6. Numerical results

The numerical section is divided in three subsections. In §4.6.1, we present the test bed that is used for all numerical experiments. In §4.6.2, we benchmark the performance of the WDT policy described in §4.5.1 against the optimal expediting policy for fixed turn-around stock. We benchmark the performance of the E-WDT heuristic against optimal joint optimization of turn-around stock and expediting policy in §4.6.3. This section also investigates how accurate the performance estimates are when L_r is approximated by an exponential distribution. Finally in 4.6.4, we present results that shed light on the value of leveraging the possibility to expedite repair in anticipating demand fluctuations. We do this in a setting with exponential lead times where the E-WDT heuristic gives optimal solutions. We compare with two naive heuristics that assume demand is a Poisson process. We also compare the E-WDT heuristic to a heuristic that determines the size of the turn-around stock and expediting policy separately.

4.6.1 Test bed and set-up

For all experiments, the backorder penalty cost is fixed at $p = 10$. The other problem parameters are varied as a full factorial experiment. The expected additional regular repair lead time was either low or high, $\mathbb{E}[L_r] \in \{2, 4\}$, and takes on an Erlang distribution, i.e., $\mu_1 = \mu_2 = \ldots = \mu_m$. The level of detail with which order progress is tracked, as modeled by m, is varied between 1 and 6, depending on what is computationally feasible. (For example the computational burden is higher when demand is a MMPP as opposed to a stationary Poisson process.) The parameter m is shown when results are presented so that it is always clear exactly how m was varied. The expedited repair lead time is either low or high, $\ell_e \in \{1, 2\}$. By Proposition 4.1, we know that expediting is only useful when $c_e < p\mathbb{E}[L_r]$. Therefore, we chose $c_e = \nu p\mathbb{E}[L_r]$ for $\nu \in \{0.2, 0.4\}$.

Demand is a stationary Poisson process or a MMPP. The long run average demand intensity $\bar{\lambda} \in \{1, 2\}$. For the Markov modulated Poisson demand, we use two basic 'modulating processes' that we refer to as the cyclic and erratic MMPP respectively. They are specified by the generator matrices and intensity vectors

$$\mathbf{Q}_{\text{cyclic}} = \cdot \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix}, \qquad \boldsymbol{\lambda}_{\text{cyclic}} = \begin{pmatrix} \frac{1}{2} & 1 & \frac{3}{2} & 1 \end{pmatrix}$$

Parameter	#	Values
Long run average demand per period ($\bar{\lambda}$)	2	1,2
Expected additional regular lead time ($\mathbb{E}[L_r]$)	2	2,4
Expedited lead time (ℓ_e)	2	1,2
Per unit expediting costs (c_e)	2	$0.2 \cdot p\mathbb{E}[L_r]$, $0.4 \cdot p\mathbb{E}[L_r]$
Holding costs per time unit (h)	2	0.5,1
Average transition rate of modulating chain ($\bar{q}$)	2	0.1, 0.05
Turn-around stock safety factor (k)	3	0,1,2
Backorder penalty costs (p)	1	10
Basic demand process type	3	Poisson, MMPP-erratic, MMPP-cyclic

Table 4.1 Description of instances in our test bed

$$\mathbf{Q}_{\text{erratic}} = \cdot \begin{pmatrix} -\frac{3}{2} & 1 & \frac{1}{2} \\ 1 & -\frac{3}{2} & \frac{1}{2} \\ \frac{2}{5} & \frac{2}{5} & -\frac{4}{5} \end{pmatrix}, \qquad \boldsymbol{\lambda}_{\text{erratic}} = \begin{pmatrix} \frac{1}{4} & \frac{1}{2} & 2 \end{pmatrix}$$

It is readily verified that both these MMPPs have a long run average demand of 1 per time unit. Therefore, by multiplying $\boldsymbol{\lambda}$ by $\bar{\lambda}$ we obtain a MMPP with a long run average demand of $\bar{\lambda}$ per time unit. Secondly, we scale how quickly the modulating chain of demand evolves by pre-multiplying the generator of $Y(t)$, $\mathbf{Q}$, by a scalar $\bar{q}$. For our experiment, $\bar{q} \in \left\{\frac{1}{20}, \frac{1}{10}\right\}$ so that the demand environment fluctuates either quickly or slowly relative to the replenishment lead times. Note that $\bar{q}$ does not affect the stationary distribution of the modulating chain of demand, and so it does not affect the long run average demand per time unit. However, it does affect the variability of demand over any finite time horizon.

In §4.6.2, we investigate the performance of the WDT policy for fixed turn-around stock. In this section, the fixed turn-around stock is set as

$$S := \left\lceil \lambda(\mathbb{E}[L_r] + \ell_e) + k\sqrt{\lambda(\mathbb{E}[L_r] + \ell_e)} \right\rceil$$

with $k \in \{0, 1, 2\}$. ($\lceil x \rceil$ denotes x rounded up to the nearest integer.) We refer to k as the safety factor and turn-around stock is tight for $k = 0$ up to ample for $k = 2$.

In §4.6.3, we optimize the expediting policy jointly with the turn-around stock. In this section, the test bed has two levels of holding cost, $h \in \left\{\frac{1}{2}, 1\right\}$.

A summary of the test bed is given in Table 4.1.

In the numerical experiments, we use value iteration to determine $C(S)$, $C_{\text{E}}(S)$, and $C_{\text{R}}(S)$. (We used Bellman equation (4.2) with the smaller uniformization constant in our algorithm). All value iteration algorithms are implemented in C and the value iteration is terminated when the relative error is less then 10^{-4}, i.e., value iteration

to find optimal policies stops after $n+1$ iterations if

$$\frac{\max_{(\mathbf{x},y)\in\mathcal{S}}(V_{n+1}(\mathbf{x},y)-V_n(\mathbf{x},y))-\min_{(\mathbf{x},y)\in\mathcal{S}}(V_{n+1}(\mathbf{x},y)-V_n(\mathbf{x},y))}{\frac{1}{2}[\max_{(\mathbf{x},y)\in\mathcal{S}}(V_{n+1}(\mathbf{x},y)-V_n(\mathbf{x},y))+\min_{(\mathbf{x},y)\in\mathcal{S}}(V_{n+1}(\mathbf{x},y)-V_n(\mathbf{x},y))]}<10^{-4}.$$

To evaluate a given WDT policy exactly when $m \neq 1$ (i.e. evaluate $C_R(S)$), we use value iteration with the same stopping criterion.

4.6.2 Performance of the WDT policy for fixed turn-around stock

In this section, we investigate how the WDT policy performs relative to the optimal expediting policy for fixed turn-around stock. The turn-around stock is fixed so that the expediting policy can be studied in isolation from the stocking decision. To this end, we investigate the relative difference of the WDT obtained by assuming L_r is exponential with respect to the optimal expediting policy. Formally this is defined as:

$$\delta_C = \frac{C_{\mathrm{E}}(S)-C(S)}{C(S)}\cdot 100\%,$$

Tables 4.2 and 4.3 show the average and maximum optimality gaps for this situation over the test bed in §4.6.1. In both cases, the optimality gaps increase as m increases, meaning that order progress information does have added value, especially for more predictable demand (Poisson demand and cyclic MMPP demand). When demand is not very predictable (erratic MMPP demand), the variance of demand increases and so do the costs. The value of order progress information remains relatively steady and so the relative value of this information decreases. However, the magnitude of all optimality gaps is small especially considering the fact that turn-around stock holding cost hS is not included in the costs. The larger gaps occur when k is large (Table 4.4), and consequently the turn-around stock is large. The reason for this is that when the turn-around stock is large, expediting is rarely necessary and backorders seldom occur so that penalty and expediting costs are small. In this situation, small absolute deviations from optimality can constitute large relative deviations.

The computation times for determining an optimal policy are in the order of a week when $m = 6$ and demand is a Poisson process on a machine with 2.4 GHz CPU and 4 GB of RAM. For $m = 6$ and MMPP demand, computation was no longer practical and so these results are missing from Tables 4.2 and 4.3. Throughout, when '-' appears in a table, it indicates that computation was not feasible for these instances.

Table 4.2 Average optimality gaps δ_C in backorder and expediting costs for fixed turn-around stocks

$\bar{q}$	NA	0.1	0.05	0.1	0.05	
m	Poisson	cyclic MMPP		erratic MMPP		AVG
2	0.52%	0.45%	0.43%	0.16%	0.13%	0.34%
3	0.92%	0.80%	0.76%	0.32%	0.26%	0.61%
4	1.22%	1.07%	1.02%	0.43%	0.36%	0.82%
5	1.45%	1.28%	1.22%	0.53%	0.44%	0.98%
6	1.64%	-	-	-	-	1.64%
AVG	1.15%	0.90%	0.86%	0.36%	0.30%	0.88%

Table 4.3 Maximum optimality gaps δ_C in backorder and expediting costs for fixed turn-around stocks

$\bar{q}$	NA	0.1	0.05	0.1	0.05	
m	Poisson	cyclic MMPP		erratic MMPP		MAX
2	1.92%	1.35%	1.62%	1.38%	0.84%	1.92%
3	2.76%	2.10%	2.39%	2.03%	1.24%	2.76%
4	3.38%	2.63%	2.91%	2.49%	1.58%	3.38%
5	3.86%	3.03%	3.29%	2.84%	1.89%	3.86%
6	4.25%	-	-	-	-	4.25%
MAX	4.25%	3.03%	3.29%	2.84%	1.89%	4.25%

Table 4.4 Average and maximum optimality gaps δ_C for different fixed turn-around stock sizes

k	0	1	2
AVG	0.42%	0.83%	0.96%
MAX	2.59%	4.25%	3.29%

4.6.3 Performance of the E-WDT heuristic

Now we consider the joint optimization of expediting policy and turn-around stock. In this situation the optimality gap is defined as

$$\delta_{C_{\text{tot}}} = \frac{C_{\text{R-tot}}(S_{\text{E}}) - C_{\text{tot}}(S^*)}{C_{\text{tot}}(S^*)} \cdot 100\%,$$

where $S_{\text{E}} = \text{argmin}_{S\in\mathbb{N}_0} C_{\text{E-tot}}(S)$ and $S^* = \text{argmin}_{S\in\mathbb{N}_0} C_{\text{tot}}(S)$. Tables 4.5-4.6 show that the average and maximum optimality gaps $\delta_{C_{\text{tot}}}$ have the same trends as in the case where optimization over S is not included. However, since the costs of holding repairables is now included, the optimality gaps are very small, never exceeding 0.76%. In all but 35 out of 480 instances, S^* and S_{E} coincide, and the absolute difference is never more than 1.

Table 4.5 Average optimality gaps $\delta_{C_{\text{tot}}}$ when optimization over S is included

$\bar{q}$	NA	0.1	0.05	0.1	0.05	
m	Poisson	cyclic MMPP		erratic MMPP		AVG
2	0.08%	0.09%	0.09%	0.08%	0.07%	0.08%
3	0.14%	0.17%	0.16%	0.16%	0.13%	0.15%
4	0.19%	0.23%	0.21%	0.21%	0.17%	0.20%
AVG	0.14%	0.17%	0.15%	0.15%	0.12%	0.15%

Table 4.6 Maximum optimality gaps $\delta_{C_{\text{tot}}}$ when optimization over S is included

$\bar{q}$	NA	0.1	0.05	0.1	0.05	
m	Poisson	cyclic MMPP		erratic MMPP		MAX
2	0.22%	0.23%	0.29%	0.39%	0.23%	0.39%
3	0.36%	0.34%	0.43%	0.63%	0.31%	0.63%
4	0.46%	0.44%	0.53%	0.76%	0.41%	0.76%
MAX	0.46%	0.44%	0.53%	0.76%	0.41%	0.76%

Recall that $C_{\text{R-tot}}(S)(S) \neq C_{\text{E-tot}}(S)$, because the function $C_{\text{E-tot}}(S)$ assumes that L_r has an exponential distribution. Now we investigate how closely $C_{\text{E-tot}}(S)$ approximates $C_{\text{R-tot}}(S)$ by looking at the relative error

$$\epsilon_{\text{E}} = \frac{C_{\text{E-tot}}(S_{\text{E}}) - C_{\text{R-tot}}(S_{\text{E}})}{C_{\text{R-tot}}(S_{\text{E}})} \cdot 100\%.$$

The relative error ϵ_{E} is always positive and the averages and maxima are shown in Tables 4.7 and 4.8. This observation can be explained by observing that the variability

of the exponential distribution is higher than that of the Erlang distribution. Since lead time variability generally degrades performance, we should expect that ϵ_{E} is generally positive.

Note that the approximation errors are larger than the optimality gaps shown in Tables 4.5 and 4.6, but still very acceptable. This is an important observation: *The insensitivity of the WDT policies with regard to the distribution of L_r is not only in performance evaluation, but even more so in policy optimality.* Furthermore, this approximation leads to a slight overestimation of the real costs which, from the managers perspective, is usually a safer deviation than an underestimation.

Table 4.7 Average error ϵ_{E} made by approximating L_r as having an exponential distribution when optimization over S is included

$\bar{q}$	0.1	0.05	0.1	0.05	
m	cyclic MMPP		erratic MMPP		AVG
2	0.46%	0.26%	0.89%	0.52%	0.53%
3	0.63%	0.35%	1.23%	0.71%	0.73%
4	0.73%	0.40%	1.42%	0.81%	0.84%
AVG	0.61%	0.33%	1.18%	0.68%	0.70%

Table 4.8 Maximum error ϵ_{E} made by approximating L_r as having an exponential distribution when optimization over S is included

$\bar{q}$	0.1	0.05	0.1	0.05	
m	cyclic MMPP		erratic MMPP		MAX
2	1.13%	0.66%	1.70%	1.01%	1.70%
3	1.57%	0.89%	2.38%	1.39%	2.38%
4	1.81%	1.02%	2.76%	1.60%	2.76%
MAX	1.81%	1.02%	2.76%	1.60%	2.76%

4.6.4 Value of anticipating demand fluctuations

In this section, we discuss three simple heuristics that either ignore the fact that demand is a Markov modulated Poisson process, or that separate the expediting policy and turn-around stock sizing decisions. Thus these heuristics fail to anticipate demand fluctuations.

In the context of repairables, the Poisson process has traditionally been used to model

demand (Muckstadt, 2005; Sherbrooke, 2004). Our experience and that of Slay and Sherbrooke (1988) indicates that this model is usually accurate for short periods of time (say up to several lead times) but is not accurate for extended periods of time as demand intensity fluctuates. This effect is captured in the present model by using the MMPP to model demand. Nevertheless, it is convenient to use the Poisson demand model as the evaluation $C_{\text{E-tot}}(S)$ can be done exactly in closed form using (4.24). Consequently, an easy heuristic is to use the Poisson demand model with demand intensity either equal to the long run average demand or, to be on the safe side, equal to the demand intensity in peak periods, $\lambda_{\max}$. We refer to these two heuristics as POIS-AVG and POIS-MAX respectively when average and peak demand intensities are used.

Another common approach is to use a traditional news-vendor inventory model without expediting to determine the turn-around stock. A simple approach would be to select S to minimize

$$hS + p\mathbb{E}\left[\left(D^Y_{t,t+\ell_e+\mathbb{E}[L_r]} - S\right)^+\right]. \tag{4.25}$$

The S that minimizes (4.25) is the smallest integer that satisfies the newsvendor inequality

$$\sum_{y\in\Theta} \mathbb{P}\left\{D^y_{t,t+\ell_e+\mathbb{E}[L_r]} \le S\right\} \mathbb{P}\{Y = y\} \ge \frac{p-h}{p}, \tag{4.26}$$

and we denote this minimizer by S_{NVF}. (NVF is short for newsvendor fractile.) After determining S_{NVF} using (4.26), we determine the optimal expediting policy for S_{NVF}. We refer to this heuristic as NVF-D$_\infty$.

We consider the case that L_r has an exponential distribution so that the optimal expediting policy is a WDT policy. The POIS-AVG and POIS-MAX coincide with the optimal and E-WDT solution when demand is a Poisson process for all the instances in our test bed. (The reason for this is that $C_{\text{tot}}(S)$ is unimodal.) Therefore, Tables 4.9 and 4.10 show the average and maximum optimality gaps only for the cases that demand is an MMPP for all three naive heuristics: POIS-AVG, POIS-MAX and NVF-D$_\infty$.

When demand is relatively steady, as is the case in the cyclic MMPP demand proces, POIS-AVG does not perform very bad, but when demand is more erratic, the performance deteriorates dramatically with optimality gaps up to 63.67%. The POIS-MAX policy avoids these extreme optimality gaps (although 24.21% is still quite substantial), but this occurs at the expense of cost performance when demand follows the more moderate cyclic MMPP process. The NVF-D$_\infty$ solution performs similarly for all demand scenario's. In general, however all naive heuristics performs quite poorly with an average optimality gap of more than 11.69% for each heuristic

Table 4.9 Average optimality gaps of naive heuristics

$\bar{q}$	0.1	0.05	0.1	0.05	
Heuristic	cyclic MMPP		erratic MMPP		AVG
POIS-AVG	2.41%	3.04%	19.48%	23.26%	12.05%
POIS-MAX	12.07%	11.32%	12.64%	11.17%	11.80%
NVF-D_∞	11.61%	10.99%	11.93%	12.26%	11.69%
AVG	8.70%	8.45%	14.68%	15.56%	11.85%

Table 4.10 Maximum optimality gaps for naive heuristics

$\bar{q}$	0.1	0.05	0.1	0.05	
Heuristic	cyclic MMPP		erratic MMPP		MAX
POIS-AVG	6.14%	8.07%	51.78%	63.67%	63.67%
POIS-MAX	15.38%	15.35%	24.21%	23.43%	24.21%
NVF-D_∞	29.44%	29.40%	29.72%	34.15%	34.15%
MAX	29.44%	29.40%	51.78%	63.67%	63.67%

and maximum optimality gaps above 24.21% for each heuristic. Consequently, we conclude that

1. There is great value in leveraging knowledge about demand fluctuations, as contained in $Y(t)$ for making repair expediting decisions.
2. Fluctuations of demand and the possibility to anticipate these through expediting repairs should be considered explicitly in sizing the turn-around stock and can lead to substantial savings.

4.7. Conclusion

In this chapter, we have considered the joint problem of finding the best turn-around stock and expediting policy for repairables that experience fluctuating demand. With regard to expediting policies, we have characterized the structure of optimal policies, confirming a conjecture by Song and Zipkin (2009). Since computing optimal expediting policies suffers from the curse of dimensionality, we proposed the use of WDT policies. These policies have an intuitive appeal and share important monotonicity properties with optimal policies.

We have shown that the joint problem can be solved by using convexity of the problem

with respect to the turn-around stock for optimal expediting policies as well as WDT policies that are obtained by assuming that the additional regular repair lead time, L_r, has an exponential distribution.

In a numerical study, we have shown that the E-WDT heuristic we propose performs very close to optimal with an optimality gap of 0.15% on average and 0.76% at most across our test bed.

Finally, we investigated the value of anticipating demand fluctuations by proper joint optimization of the turn-around stock and expediting policy by comparing the E-WDT heuristic with more naive heuristics that do not anticipate demand fluctuations or that separate the stocking and expediting problems. With optimality gaps of 11.85% on average and of at most 63.67%, we have shown that

1. There is great value in leveraging knowledge about demand fluctuations, when making repair expediting decisions.
2. Fluctuations of demand and the possibility to anticipate these through expediting repairs should be considered explicitly in sizing the turn-around stock and can lead to substantial savings.

4.A. Determining $\mathbb{P}\{D^y_{t,t+\ell_e} = k\}$

In this appendix, we show how $\mathbb{P}\left\{D^y_{t,t+\ell_e} = k\right\}$ can be determined numerically. To this end, let $p_{y,y'}(k,\ell_e) = \mathbb{P}\left\{D^y_{t,t+\ell_e} = k | Y(t+\ell_e) = y'\right\}$ be the (y,y')-entry of the matrix $\mathbf{P}(k,\ell_e)$. Then the matrix generating function $\widetilde{\mathbf{P}}(z,\ell_e) = \sum_{k=0}^{\infty} \mathbf{P}(k,\ell_e) z^k$ satisfies (e.g Fischer and Meier-Hellstern, 1992):

$$\widetilde{\mathbf{P}}(z,\ell_e) = \exp\left([\mathbf{Q} - (1-z)\,\mathrm{diag}(\boldsymbol{\lambda})]\ell_e\right).$$

A plethora of numerical methods to compute the matrix exponential are discussed in Moler and Van Loan (2003). For the numerical work in this thesis, we use the scaling and squaring algorithm with a Padé approximation. The probabilities $\mathbb{P}\left\{D^y_{t,t+\ell_e} = k | Y(t+\ell_e) = y'\right\}$ can be obtained from $\widetilde{\mathbf{P}}(z,\ell_e)$ by numerical inversion using the LATTICE-POISSON algorithm of Abate and Whitt (1992) which uses the approximation

$$\begin{aligned}
&\mathbb{P}\left\{D^y_{t,t+\ell_e} = k | Y(t+\ell_e) = y'\right\} \\
&\quad \approx \frac{1}{2kr^k}\left\{\widetilde{\mathbf{P}}(r,\ell_e) + (-1)^k \widetilde{\mathbf{P}}(-r,\ell_e) + 2\sum_{n=1}^{k-1}(-1)^n \,\mathrm{Re}(\widetilde{\mathbf{P}}(r\exp(n\pi \mathrm{i}/k),\ell_e))\right\},
\end{aligned}$$

where $\mathrm{i} = \sqrt{-1}$, $0 < r < 1$ and $\mathrm{Re}(x)$ denotes the real part of the complex number x. The absolute error in this approximation is bounded by $\frac{r^{2k}}{1-r^{2k}}$ and so by choosing $r = 10^{-\gamma/(2k)}$, we obtain an accuracy of approximately $10^{-\gamma}$. Then the needed probability, $\mathbb{P}\left\{D^y_{t,t+\ell_e} = k\right\}$, can be found by un-conditioning:

$$\mathbb{P}\left\{D^y_{t,t+\ell_e} = k\right\} = \sum_{y' \in \Theta} \mathbb{P}\left\{D^y_{t,t+\ell_e} = k | Y(t+\ell_e) = y'\right\} \mathbb{P}\{Y(t+\ell_e) = y' | Y(t) = y\}$$

The probabilities $\mathbb{P}\{Y(t+\ell_e) = y' | Y(t) = y\}$ are found from the transient analysis of $Y(t)$. In particular, if we let $r_{y,y'} = \mathbb{P}\{Y(t+\ell_e) = y' | Y(t) = y\}$ be the (y,y')-th element of the matrix $\mathbf{R}(\ell_e)$, then $\mathbf{R}(\ell_e) = \exp(\ell_e \mathbf{Q})$.

4.B. Proofs

4.B.1 Proof of Lemma 4.1

PROOF: The proof is a direct proof. Note that $\mathbb{P}\left\{D^y_{t,t+\ell_e} = k\right\} = 0$ for $k < 0$. For part (i) we have:

$$\begin{aligned}
\Delta c_p(x,y) &= c_p(x+1,y) - c_p(x,y) \\
&= p\left[\sum_{k=S-(x+1)}^{\infty}(k-S+x+1)\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\}\right. \\
&\quad \left. - \sum_{k=S-x}^{\infty}(k-S+x)\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\}\right] \\
&= p\left[\sum_{k=S-x-1}^{\infty}(k-S+x)\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\} + \sum_{k=S-x-1}^{\infty}\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\}\right. \\
&\quad \left. - \sum_{k=S-x}^{\infty}(k-S+x)\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\}\right] \\
&= p\sum_{k=S-x}^{\infty}\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\} \geq 0. \qquad (4.27)
\end{aligned}$$

For part (ii) we have:

$$\begin{aligned}
\Delta^2 c_p(x,y) &= \Delta c_p(x+1,y) - \Delta c_p(x,y) \\
&= p\left[\sum_{k=S-(x+1)}^{\infty}\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\} - \sum_{k=S-x}^{\infty}\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\}\right] \\
&= p\mathbb{P}\left\{D^y_{t,t+\ell_e}=S-x-1\right\} \geq 0. \qquad (4.28)
\end{aligned}$$

Part (iii) follows immediately from (4.27) and noting that $\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\} = 0$ for $k < 0$.

Finally for part (iv), we can write using (4.27)

$$\begin{aligned}
&\Delta c_p(x,y|S) - \Delta c_p(x,y|S+1) \\
&\quad = p\left[\sum_{k=S-x}^{\infty}\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\} - \sum_{k=S+1-x}^{\infty}\mathbb{P}\left\{D^y_{t,t+\ell_e}=k\right\}\right] \\
&\quad = p\mathbb{P}\left\{D^y_{t,t+\ell_e}=S-x\right\} \geq 0 \qquad (4.29)
\end{aligned}$$

□

4.B.2 Proof of Proposition 4.1 (ii)

PROOF: Here we present the proof of part (ii), the proof of part (i) is in the main text. The proof is based on constructing two coupled processes and showing that

expediting repair is expected to dominate using regular repair when there are many parts still undergoing additional regular repair.

Let $\varepsilon = (p\mathbb{E}[L_r] - c_e)/3 > 0$. We denote the probability density function of L_r as f_{L_r} and fix $\alpha < \infty$ to verify

$$\int_{t=\alpha}^{\infty} t f_{L_r}(t_r) dt \leq \varepsilon/p. \tag{4.30}$$

Such an α exists because $tf_{L_r}(t) > 0$ for $t \in (0, \infty)$ so that $\int_{t=\alpha}^{\infty} tf_{L_r}(t)dt$ is strictly decreasing in α and furthermore $\lim_{\alpha\to\infty} \int_{t=\alpha}^{\infty} tf_{L_r}(t)dt = 0$. Let E_μ denote an exponential random variable with mean μ^{-1}. We fix an integer M' to verify

$$\mathbb{P}\left\{E_{\mu_m} < \alpha\right\}^{M'} \mathbb{E}[L_r] \leq \varepsilon. \tag{4.31}$$

Such an $M' \in \mathbb{N}$ exists because $\alpha < \infty$ and so $\mathbb{P}\left\{E_{\mu_m} < \alpha\right\} < 1$.

Now we consider an arbitrary policy π that does *not* expedite when $\mathbf{x}\mathbf{e}^{\mathrm{T}} \geq S+M' = M$ for some $(\mathbf{x}, y) \in \mathcal{S}$. Consider an arbitrary moment in time, t', when a failed part arrives to the system and $\sum_{i=1}^{m} X_i(t') \geq S + M' = M$ and policy π stipulates that the part should *not* be expedited. Denote this process $\mathbf{X}^\pi(t)$. We let $\widetilde{\mathbf{X}}(t)$ denote the evolution of the part sent to regular repair at time t' by policy π, so $\widetilde{\mathbf{X}}(t) = \mathbf{e}_i$ if the part sent to repair at time t' has completed its first $i-1$ phases of the additional regular repair, and $\widetilde{\mathbf{X}}(t) = \mathbf{0}$ if the part has completed its additional regular repair lead time. Next, we consider an alternate process which is identical to $\mathbf{X}^\pi(t)$ except that it does expedite the unit arriving at t'. We denote this process $\mathbf{X}^e(t)$, and formally define it as $\mathbf{X}^e(t) = \mathbf{X}^\pi(t) - \widetilde{\mathbf{X}}(t)$. We let $T_r = \inf\{t - t' | \widetilde{\mathbf{X}}(t) = \mathbf{0}, t \geq t'\}$ and note that $T_r \overset{d}{=} L_r$.

Analogous to the proof of part (i), $\mathbf{X}^\pi(t)\mathbf{e}^{\mathrm{T}} = \mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}} + 1$ for $t \in [t', t' + T_r)$, and $\mathbf{X}^\pi(t) = \mathbf{X}^e(t)$ for $t \geq t' + T_r$. Also both processes make exactly the same expediting decisions for all $t > t'$. Thus any cost differences between $\mathbf{X}^e(t)$ and $\mathbf{X}^\pi(t)$ occur in the time interval $[t', t' + T_r)$. Denote the expectation of this cost difference Ξ. Then we have:

$$\begin{aligned}
\Xi &= c_e - \mathbb{E}_{T_r}\left\{\mathbb{E}_{(\mathbf{X}^e(t), Y(t))}\left[\int_{t=t'}^{t'+T_r} \Delta c_p(\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}}, Y(t))dt \,\middle|\, T_r\right]\right\} \\
&= c_e - \int_{t_r=0}^{\infty} \mathbb{E}_{(\mathbf{X}^e(t), Y(t))}\left[\int_{t=t'}^{t'+T_r} \Delta c_p(\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}}, Y(t))dt \,\middle|\, T_r = t_r\right] f_{L_r}(t_r)dt_r \\
&= c_e - \int_{t_r=0}^{\infty} \mathbb{E}_{(\mathbf{X}^e(t), Y(t))}\left[\int_{t=t'}^{t'+T_r} \Delta c_p(\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}}, Y(t))dt \,\middle|\, T_r = t_r,\, \mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}} \geq S \text{ for all } t \in (t', t' + T_r)\right] \\
&\quad \times \mathbb{P}\{\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}} \geq S \text{ for all } t \in (t', t' + t_r)\} f_{L_r}(t_r)dt_r \\
&\quad - \int_{t_r=0}^{\infty} \mathbb{E}_{(\mathbf{X}^e(t), Y(t))}\left[\int_{t=t'}^{t'+T_r} \Delta c_p(\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}}, Y(t))dt \,\middle|\, T_r = t_r,\, \mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}} < S \text{ for some } t \in (t', t' + T_r)\right]
\end{aligned}$$

$$
\begin{aligned}
&\times \mathbb{P}\{\mathbf{X}_e(t)\mathbf{e}^{\mathrm{T}} < S \text{ for some } t \in (t', t'+t_r)\} f_{L_r}(t_r) dt_r \\
&\leq c_e - \sum_{t_r=0}^{\infty} \mathbb{E}_{(\mathbf{X}^e(t), Y(t))} \left[\int_{t=t'}^{t'+T_r} \Delta c_p(\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}}, Y(t)) dt \,\middle|\, T_r = t_r,\ \mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}} \geq S \text{ for all } t \in (t', t'+T_r) \right] \\
&\quad \times \mathbb{P}\{\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}} \geq S \text{ for all } t \in (t', t'+t_r)\} f_{L_r}(t_r) dt_r \\
&= c_e - \sum_{t_r=0}^{\infty} p t_r f_{L_r}(t_r) \mathbb{P}\{\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}} \geq S \text{ for all } t \in (t', t'+t_r)\} dt_r \qquad (4.32)
\end{aligned}
$$

The third equality is obtained by conditioning on whether or not $\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}}$ stays above S on the interval $[t', t'+T_r)$. The first inequality follows from dropping the last term and the last equality follows from Lemma 4.1 (iii).

Next we observe that $\mathbb{P}\{\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}} \geq S \text{ for all } t \in (t', t'+t_r)\}$ is bounded below by the probability that there are fewer than M' parts already in additional regular repair at time t', finish additional regular repair before $t'+t_r$. Since the remaining time in regular repair for any of these parts is at least an E_{μ_m} random variable (by the lack of memory property), we conclude that

$$
\mathbb{P}\{\mathbf{X}^e(t)\mathbf{e}^{\mathrm{T}} \geq S \text{ for all } t \in (t', t'+t_r)\} \geq 1 - \mathbb{P}\{E_{\mu_m} < t_r\}^{M'}. \qquad (4.33)
$$

Now continuing from (4.32) and using (4.33) we obtain:

$$
\begin{aligned}
\Xi &\leq c_e - \int_{t_r=0}^{\infty} p t_r f_{L_r}(t_r) \left(1 - \mathbb{P}\{E_{\mu_m} < t_r\}^{M'}\right) dt_r \\
&= c_e - p\mathbb{E}[L_r] + \int_{t_r=0}^{\infty} p t_r f_{L_r}(t_r) \mathbb{P}\{E_{\mu_m} < t_r\}^{M'} dt_r \\
&= c_e - p\mathbb{E}[L_r] + \int_{t_r=0}^{\alpha} p t_r f_{L_r}(t_r) \mathbb{P}\{E_{\mu_m} < t_r\}^{M'} dt_r \\
&\quad + \int_{t_r=\alpha}^{\infty} p t_r f_{L_r}(t_r) \mathbb{P}\{E_{\mu_m} < t_r\}^{M'} dt_r \\
&\leq c_e - p\mathbb{E}[L_r] + \mathbb{P}\{E_{\mu_m} < \alpha\}^{M'} \int_{t_r=0}^{\alpha} p t_r f_{L_r}(t_r) dt_r + \int_{t_r=\alpha}^{\infty} p t_r f_{L_r}(t_r) dt_r \qquad (4.34) \\
&\leq c_e - p\mathbb{E}[L_r] + \mathbb{P}\{E_{\mu_m} < \alpha\}^{M'} \mathbb{E}[L_r] + \int_{t_r=\alpha}^{\infty} p t_r f_{L_r}(t_r) dt_r \\
&\leq -3\varepsilon + \varepsilon + \varepsilon = -\varepsilon < 0. \qquad (4.35)
\end{aligned}
$$

Inequality (4.34) follows because $\mathbb{P}\{E_{\mu_m} < t_r\}$ is increasing in t_r and the final inequalities follow from the choice of ε, α and M'. Since $\Xi < 0$, we conclude that the expected cost of process $\mathbf{X}^{\pi}(t)$ is greater than the cost of $\mathbf{X}^e(t)$. Thus, we have shown that any policy that does not expedite when $\mathbf{X}(t)\mathbf{e}^{\mathrm{T}} \geq M$ and $c_e < p\mathbb{E}[L_r]$ can be strictly improved by expediting whenever $\mathbf{X}(t)\mathbf{e}^{\mathrm{T}} \geq M$. That is, if $c_e < p\mathbb{E}[L_r]$, then there is a $M \in \mathbb{N}$ such that whenever $\mathbf{X}(t)\mathbf{e}^{\mathrm{T}} \geq M$ it is optimal to expedite. □

4.B.3 Proof of Lemma 4.2

To facilitate the presentation of the proof we introduce the following shorthand:

$$\mathcal{I} = \bigcap_{i=1}^{m} \mathcal{I}(i), \quad \mathcal{SM}(1) = \bigcap_{j=2}^{m} \mathcal{SM}(1,j)$$

Furthermore, when $f \in \mathcal{X}$ implies $\mathbb{T}_{\mathcal{Y}} f(\mathbf{x}, y) \in \mathcal{X}$, we say that $\mathbb{T}_{\mathcal{Y}}$ propagates $\mathcal{X}$.

PROOF: Part (i) of the lemma can be verified directly by using Lemma 4.1.

For part (ii) we consider each operator separately. Let $f \in \mathcal{F}$. For operator $\mathbb{T}_{\text{cost}}$ the results hold because of part (i) of this lemma and Theorem 7.1 in Koole (2006). For $\mathbb{T}_{\text{env}}$ the result hold trivially as this operator produces linear combinations of functions in $\mathcal{F}$.

By Theorems 7.3 and 7.4 of Koole (2006) we have that $\mathbb{T}_{\text{TD}(i)}$ propagates $\mathcal{F}$ for $i = 1, ..., m-1$ and $\mathbb{T}_{\text{D}(m)}$ propagates $\mathcal{F}$.

For $\mathbb{T}_{\text{AC}(1)}$, the inequalities that characterize $\mathcal{I} \cap \mathcal{UI}$ are propagated whenever $\mathbf{x}\mathbf{e}^{\mathrm{T}} < M-1$ by Theorem 7.2 of Koole (2006). When $\mathbf{x}\mathbf{e}^{\mathrm{T}} = M-1$ we have for $i = 1, \ldots, m$:

$$\begin{aligned} \Delta_i \mathbb{T}_{\text{AC}(1)} f(\mathbf{x}, y) &= c_e + f(\mathbf{x} + \mathbf{e}_i, y) - \min(c_e + f(\mathbf{x}, y), f(\mathbf{x} + \mathbf{e}_1, y)) \\ &\geq c_e + f(\mathbf{x} + \mathbf{e}_i, y) - c_e - f(\mathbf{x}, y) \\ &= f(\mathbf{x} + \mathbf{e}_i, y) - f(\mathbf{x}, y) \geq 0, \end{aligned} \tag{4.36}$$

where the second inequality holds because $f \in \mathcal{I}$. This shows $\mathbb{T}_{\text{AC}(1)}$ propagates $\mathcal{I}$. Similarly, and again for $\mathbf{x}\mathbf{e}^{\mathrm{T}} = M-1$, we find for $i = 1, \ldots, m-1$:

$$\mathbb{T}_{\text{AC}(1)} f(\mathbf{x} + \mathbf{e}_i, y) - \mathbb{T}_{\text{AC}(1)} f(\mathbf{x} + \mathbf{e}_{i+1}, y) = c_e + f(\mathbf{x} + \mathbf{e}_i, y) - c_e - f(\mathbf{x} + \mathbf{e}_{i+1}, y) \geq 0, \tag{4.37}$$

where the inequality holds because $f \in \mathcal{UI}$. Thus we have shown that $\mathbb{T}_{\text{AC}(1)}$ propagates $\mathcal{UI}$. (Recall that the case $\mathbf{x}\mathbf{e}^{\mathrm{T}} = M$ need not be considered because, in this case, the inequalities do not exist in $\mathcal{S}$. A similar observation will hold for the other inequalities in $\mathcal{F}$.) Also by Theorem 7.2 of Koole (2006), for all $\mathbf{x}$ that satisfy $\mathbf{x}\mathbf{e}^{\mathrm{T}} < M-2$ it holds that $\mathbb{T}_{\text{AC}(1)} f(\mathbf{x}, y) \in \mathcal{C}(1) \cap \mathcal{SM}(1)$. Consider the case that $\mathbf{x}\mathbf{e}^{\mathrm{T}} = M-2$. To show $\mathbb{T}_{\text{AC}(1)}$ preserves convexity, we consider three cases:

(a) Case: $\min\{c_e + f(\mathbf{x} + \mathbf{e}_1, y), f(\mathbf{x} + 2\mathbf{e}_1, y)\} = f(\mathbf{x} + 2\mathbf{e}_1, y)$. This case implies that $c_e \geq f(\mathbf{x} + 2\mathbf{e}_1, y) - f(\mathbf{x} + \mathbf{e}_1, y)$ and furthermore as $f \in \mathcal{C}(1)$ we have $\min\{c_e + f(\mathbf{x}, y), f(\mathbf{x} + \mathbf{e}_1, y)\} = f(\mathbf{x} + \mathbf{e}_1, y)$. Thus we have:

$$\begin{aligned} \Delta_1^2 \mathbb{T}_{\text{AC}(1)} f(\mathbf{x}, y) &= c_e + f(\mathbf{x} + 2\mathbf{e}_1, y) - 2f(\mathbf{x} + 2\mathbf{e}_1, y) + f(\mathbf{x} + \mathbf{e}_1, y) \\ &= c_e - f(\mathbf{x} + 2\mathbf{e}_1, y) + f(\mathbf{x} + \mathbf{e}_1, y) \geq 0. \end{aligned} \tag{4.38}$$

The inequality holds because $c_e \geq f(\mathbf{x} + 2\mathbf{e}_1, y) - f(\mathbf{x} + \mathbf{e}_1, y)$.

(b) Case: $\min\{c_e + f(\mathbf{x}+\mathbf{e}_1, y), f(\mathbf{x}+2\mathbf{e}_1, y)\} = c_e + f(\mathbf{x}+\mathbf{e}_1, y)$ and $\min\{c_e + f(\mathbf{x}, y), f(\mathbf{x}+\mathbf{e}_1, y)\} = c_e + f(\mathbf{x}, y)$. Now we have

$$\begin{aligned}\Delta_1^2 \mathbb{T}_{\mathrm{AC}(1)} f(\mathbf{x}, y) &= c_e + f(\mathbf{x}+2\mathbf{e}_1, y) - 2c_e - 2f(\mathbf{x}+\mathbf{e}_1, y) + c_e + f(\mathbf{x}, y) \\ &= f(\mathbf{x}+2\mathbf{e}_1, y) - 2f(\mathbf{x}+\mathbf{e}_1, y) + f(\mathbf{x}, y) \geq 0, \end{aligned} \tag{4.39}$$

where the inequality holds because $f \in \mathcal{C}(1)$.

(c) Case: $\min\{c_e + f(\mathbf{x}+\mathbf{e}_1, y), f(\mathbf{x}+2\mathbf{e}_1, y)\} = c_e + f(\mathbf{x}+\mathbf{e}_1, y)$ and $\min\{c_e + f(\mathbf{x}, y), f(\mathbf{x}+\mathbf{e}_1, y)\} = f(\mathbf{x}+\mathbf{e}_1, y)$. Now we have:

$$\begin{aligned}\Delta_1^2 \mathbb{T}_{\mathrm{AC}(1)} f(\mathbf{x}, y) &= c_e + f(\mathbf{x}+2\mathbf{e}_1, y) - 2c_e - 2f(\mathbf{x}+\mathbf{e}_1) + f(\mathbf{x}+\mathbf{e}_1, y) \\ &= f(\mathbf{x}+2\mathbf{e}_1, y) - f(\mathbf{x}+\mathbf{e}_1, y) - c_e \geq 0. \end{aligned} \tag{4.40}$$

The inequality holds because the case implies that $c_e \leq f(\mathbf{x}+2\mathbf{e}_1, y) - f(\mathbf{x}+\mathbf{e}_1, y)$.

Thus we have shown that $\mathbb{T}_{\mathrm{AC}(1)}$ propagates $\mathcal{C}(1)$ if $f \in \mathcal{F}$. To show $\mathbb{T}_{\mathrm{AC}(1)}$ also propagates $\mathcal{SM}(1)$, we distinguish 2 cases.

(a) Case: $\min\{c_e + f(\mathbf{x}, y), f(\mathbf{x}+\mathbf{e}_1, y)\} = c_e + f(\mathbf{x}, y)$. In this case we have

$$\begin{aligned}&\Delta_1 \mathbb{T}_{\mathrm{AC}(1)} f(\mathbf{x}+\mathbf{e}_j, y) - \Delta_1 \mathbb{T}_{\mathrm{AC}(1)} f(\mathbf{x}, y) \\ &\quad = c_e + f(\mathbf{x}+\mathbf{e}_1+\mathbf{e}_j, y) - \min\{c_e + f(\mathbf{x}+\mathbf{e}_j, y), f(\mathbf{x}+\mathbf{e}_1+\mathbf{e}_j, y)\} \\ &\qquad - \min\{c_e + f(\mathbf{x}+\mathbf{e}_1, y), f(\mathbf{x}+2\mathbf{e}_1)\} + \min\{c_e + f(\mathbf{x}, y), f(\mathbf{x}+\mathbf{e}_1, y)\} \\ &\quad \geq 2c_e + f(\mathbf{x}+\mathbf{e}_1+\mathbf{e}_j) - 2c_e - f(\mathbf{x}+\mathbf{e}_j, y) - f(\mathbf{x}+\mathbf{e}_1, y) + f(\mathbf{x}, y) \\ &\quad = f(\mathbf{x}+\mathbf{e}_1+\mathbf{e}_j) - f(\mathbf{x}+\mathbf{e}_j, y) - f(\mathbf{x}+\mathbf{e}_1, y) + f(\mathbf{x}, y) \geq 0. \end{aligned} \tag{4.41}$$

The second inequality holds because $f \in \mathcal{SM}(1, j)$.

(b) Case: $\min(c_e + f(\mathbf{x}, y), f(\mathbf{x}+\mathbf{e}_1, y)) = f(\mathbf{x}+\mathbf{e}_1, y)$ Now we find that:

$$\begin{aligned}&\Delta_1 \mathbb{T}_{\mathrm{AC}(1)} f(\mathbf{x}+\mathbf{e}_j, y) - \Delta_1 \mathbb{T}_{\mathrm{AC}(1)} f(\mathbf{x}, y) \\ &\quad = c_e + f(\mathbf{x}+\mathbf{e}_1+\mathbf{e}_j, y) - \min\{c_e + f(\mathbf{x}+\mathbf{e}_j, y), f(\mathbf{x}+\mathbf{e}_1+\mathbf{e}_j, y)\} \\ &\qquad - \min\{c_e + f(\mathbf{x}+\mathbf{e}_1, y), f(\mathbf{x}+2\mathbf{e}_1)\} + \min\{c_e + f(\mathbf{x}, y), f(\mathbf{x}+\mathbf{e}_1, y)\} \\ &\quad \geq c_e + f(\mathbf{x}+\mathbf{e}_1+\mathbf{e}_j, y) - f(\mathbf{x}+\mathbf{e}_1+\mathbf{e}_j, y) \\ &\qquad - c_e - f(\mathbf{x}+\mathbf{e}_1, y) + f(\mathbf{x}+\mathbf{e}_1, y) = 0. \end{aligned} \tag{4.42}$$

Thus we have shown that $\mathbb{T}_{\mathrm{AC}(1)}$ propagates $\mathcal{SM}(1)$ if $f \in \mathcal{F}$.

Part (iii) holds trivially as $\mathbb{T}_{\mathrm{unif}}$ produces linear combinations of functions in $\mathcal{F}$. □

4.B.4 Proof of Lemma 4.4

PROOF: The property in (4.23) is a type of submodularity property (c.f. Altman and Koole, 1998). In this proof, we actually prove a slightly stronger property, namely that (4.23) also holds with Δ_1 replaced by Δ_i for $i = 1, \ldots, m$. We define $\mathcal{SB}$ as:

$$\mathcal{SB} = \{f \in \mathcal{W} | \Delta_i f(\mathbf{x}, y, S) \geq \Delta_i f(\mathbf{x}, y, S+1) \text{ for } i = 1, \ldots, m\}.$$

Because of (4.12) and the fact that $V_0(\mathbf{x}, y, S) \in \mathcal{SB}$, we only need to show that if $f \in \mathcal{SB}$ and $f_j \in \mathcal{SB}$ for $j = 1, \ldots, m+2$, then also

$$\begin{aligned}
&\mathbb{T}_{\text{cost}} f(\mathbf{x}, y, S), \mathbb{T}_{\text{AC}(1)} f(\mathbf{x}, y, S) \mathbb{T}_{\text{D}(m)} f(\mathbf{x}, y, S), \mathbb{T}_{\text{env}} f(\mathbf{x}, y, S) \in \mathcal{SB} \\
&\mathbb{T}_{\text{TD}(j)} f(\mathbf{x}, y, S) \in \mathcal{SB} \text{ for } j = 1, \ldots, m-1 \text{ and} \\
&\mathbb{T}_{\text{unif}}(f_1, \ldots f_{m+2})(\mathbf{x}, y, S) \in \mathcal{SB}.
\end{aligned}$$

For $\mathbb{T}_{\text{cost}}$, this follows from Lemma 4.1 (iv) and Theorem 7.1 of Koole (2006). For $\mathbb{T}_{\text{env}}$ and $\mathbb{T}_{\text{unif}}$ this follows because these operators take linear combinations of functions in $\mathcal{SB}$. For $\mathbb{T}_{\text{AC}(1)}$ and $\mathbb{T}_{\text{D}(m)}$ this follows from Theorems 7.2 and 7.3 of Koole (2006) respectively. For $\mathbb{T}_{\text{TD}(j)}$, we distinguish two cases:

(a) Case: $j \neq i$. We have

$$\begin{aligned}
&\Delta_i \mathbb{T}_{\text{TD}(j)} f(\mathbf{x}, y, S) - \Delta_i \mathbb{T}_{\text{TD}(j)} f(\mathbf{x}, y, S+1) \\
&\quad = \frac{x_j}{M} f(\mathbf{x} + \mathbf{e}_i - \mathbf{e}_j + \mathbf{e}_{j+1}, y, S) + \frac{M - x_j}{M} f(\mathbf{x} + \mathbf{e}_i, y, S) \\
&\qquad - \frac{x_j}{M} f(\mathbf{x} - \mathbf{e}_j + \mathbf{e}_{j+1}, y, S) - \frac{M - x_j}{M} f(\mathbf{x}, y, S) \\
&\qquad - \frac{x_j}{M} f(\mathbf{x} + \mathbf{e}_i - \mathbf{e}_j + \mathbf{e}_{j+1}, y, S+1) - \frac{M - x_j}{M} f(\mathbf{x} + \mathbf{e}_i, y, S+1) \\
&\qquad + \frac{x_j}{M} f(\mathbf{x} - \mathbf{e}_j + \mathbf{e}_{j+1}, y, S+1) + \frac{M - x_j}{M} f(\mathbf{x}, y, S+1) \\
&\quad = \frac{x_j}{M} \left(\Delta_i f(\mathbf{x} - \mathbf{e}_j + \mathbf{e}_{j+1}, y, S) - \Delta_i f(\mathbf{x} - \mathbf{e}_j + \mathbf{e}_{j+1}, y, S+1)\right) \\
&\qquad \frac{M - x_j}{M} \left(\Delta_i f(\mathbf{x}, y, S) - \Delta_i f(\mathbf{x}, y, S+1)\right) \geq 0
\end{aligned}$$

The inequality holds because $f \in \mathcal{SB}$.

(b) Case $j = i$. We have

$$
\begin{aligned}
&\Delta_i \mathbb{T}_{\mathrm{TD}(i)} f(\mathbf{x}, y, S) - \Delta_i \mathbb{T}_{\mathrm{TD}(i)} f(\mathbf{x}, y, S+1) \\
&= \frac{x_i + 1}{M} f(\mathbf{x} + \mathbf{e}_i - \mathbf{e}_i + \mathbf{e}_{i+1}, y, S) + \frac{M - x_i - 1}{M} f(\mathbf{x} + \mathbf{e}_i, y, S) \\
&\quad - \frac{x_i}{M} f(\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}, y, S) - \frac{M - x_i}{M} f(\mathbf{x}, y, S) \\
&\quad - \frac{x_i + 1}{M} f(\mathbf{x} + \mathbf{e}_i - \mathbf{e}_i + \mathbf{e}_{i+1}, y, S+1) - \frac{M - x_i - 1}{M} f(\mathbf{x} + \mathbf{e}_i, y, S+1) \\
&\quad + \frac{x_i}{M} f(\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}, y, S+1) + \frac{M - x_i}{M} f(\mathbf{x}, y, S+1) \\
&\geq \frac{x_i}{M} f(\mathbf{x} + \mathbf{e}_{i+1}, y, S) + \frac{M - x_i - 1}{M} f(\mathbf{x} + \mathbf{e}_i, y, S) \\
&\quad - \frac{x_i}{M} f(\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}, y, S) - \frac{M - x_i - 1}{M} f(\mathbf{x}, y, S) \\
&\quad - \frac{x_i}{M} f(\mathbf{x} + \mathbf{e}_{i+1}, y, S+1) - \frac{M - x_i - 1}{M} f(\mathbf{x} + \mathbf{e}_i, y, S+1) \\
&\quad + \frac{x_i}{M} f(\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}, y, S+1) + \frac{M - x_i - 1}{M} f(\mathbf{x}, y, S+1) \\
&= \frac{x_i}{M} \left(\Delta_i f(\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}, y, S) - \Delta_i f(\mathbf{x} - \mathbf{e}_i + \mathbf{e}_{i+1}, y, S+1) \right) \\
&\quad \frac{M - x_i - 1}{M} \left(\Delta_i f(\mathbf{x}, y, S) - \Delta_i f(\mathbf{x}, y, S+1) \right) \geq 0
\end{aligned}
$$

The first inequality follows by adding $\frac{1}{M}(\Delta_{i+1} f(\mathbf{x}, y, S+1) - \Delta_{i+1} f(\mathbf{x}, y, S))$ which is less than 0 because $f \in \mathcal{SB}$. (Note that $\Delta_{i+1} f(\mathbf{x}, y, S)$ is well defined here because because $j < m$ and so $i+1 \leq m$ because $i = j$ by assumption.) The final inequality also follows from the induction hypothesis.

□

Chapter 5

A system approach to repairable stocking and expediting in a fluctuating demand environment

"Be Prepared... the meaning of the motto is that a scout must prepare himself by previous thinking out and practicing how to act on any accident or emergency so that he is never taken by surprise."

Robert Baden-Powell

5.1. Introduction

In the previous chapter, we considered two decisions for a single repairable item: How many spare parts to buy and when to expedite the repair of a part. Although this problem can appear for a single part in isolation, it appears more commonly for one or more groups of repairables that support different fleets of equipment. Each fleet has a target with respect to availability which is translated into a requirement on the expected number of backorders for parts that belong to a fleet. Repairable spare parts are kept on stock to meet these targets for each fleet. However, this stocking problem can not be resolved for each fleet separately because repairables that belong

to different fleets may use the same resources for repair. These repair resources are flexible and this is modeled through the possibility to expedite the repair of a part. (An expedited repair order has a shorter lead time.) Since the flexibility of a repair resource is limited, there is a constraint on the amount of repair work that can be expedited per time unit for each repair resource. We refer to the amount of work that a repair resource handles per time unit as the load. Repairables from different fleets compete for the opportunity to load a repair resource with expedited orders.

For the situation described in the previous paragraph, decision makers need to determine adequate stock levels for all repairables as well as expediting rules. The objective of the decision maker is to minimize the costs involved with either purchasing or holding repairable spare parts while:

- meeting a service level in the form of a maximum average number of backorders for each fleet, and
- keeping the load imposed on each repair resource due to expedited orders below a set target level.

Demand for a single type of repairable spare part usually fluctuates over time. These demand fluctuations arise for several reasons such as periodic inspections, usage patterns of equipment over time and the season of year. (A more thorough discussion on repairable demand fluctuations over time is provided in §4.3.1.) When the reasons for demand fluctuations are understood, the expediting decision can be made to anticipate these fluctuations and to make effective use of repair resources.

In this chapter, we provide a mathematical model for the decision problem described above. This model (like the models in other chapters) has been conceived with an application at NedTrain in mind. We emphasize however that the applicability of the model and results in this chapter extend to other companies that maintain their own equipment. We will illustrate the need, as well as the application of the model using an example that runs throughout this entire chapter. This example is about a fictitious railway company. We finish this introduction by starting this example. The rest of the chapter is organized as follows. §5.2 reviews related literature and positions the contribution of this chapter with respect to existing literature. The mathematical model is provided in §5.3. The analysis of the model is in §5.4 and can be skipped without loss of continuity. Computational results of the model are provided in §5.5 and concluding remarks are offered in §5.6.

Example 5.1 The railway company Thomas&Co needs new trains to replace locomotives with pulled carriages. They decide to buy 100 trains from Liam Engineering Inc., and plan to use those for the next 30-40 years on long distance train services. Along with this order of 100 trains, Liam Engineering Inc. offers

the possibility to buy (repairable) spare parts at a considerable discounted price. Thomas&Co would like to buy repairable spare parts at this discounted price and is taking this opportunity to decide on the stocking levels of repairables for the new fleet, as well as to reconsider the stocking levels for repairables of other fleets. ◇

5.2. Literature review and contribution

Multi-item repairable inventory models are abundant in literature. We refer the reader to the books of Sherbrooke (2004) and Muckstadt (2005) and review papers by GuideSrivastava1997, Kennedy et al. (2002), and Basten and Van Houtum (2013) for a broad overview. In this section, we briefly discuss literature with similar modeling assumptions and literature that expounds on or uses similar solution methods as those used in this chapter. On the modeling side, the main contributions of this chapter are the fluctuating demand model and the use of a dynamic expediting policy that depends on demand fluctuations. On the analysis side, we decompose the problem per item via a column generation algorithm. Therefore, this section is organized around three main topics: fluctuating demand (§5.2.1), repair expediting and scheduling policies (§5.2.2), and decomposition and column generation algorithms (§5.2.3).

5.2.1 Fluctuating demand

Demand for repairables that fluctuates over time has been considered before in a series of models developed by the RAND corporation under the name Dyna-METRIC (Hillestad, 1982; Carillo, 1989; Isaacson and Boren, 1993). Initially, these models were based on an extension of Palm's theorem for non-stationary Poisson processes, but these efforts eventually developed into simulation models that do not allow efficient optimization. In the Dyna-METRIC approach, demand is a non-stationary Poisson proces, but the Poisson demand rate is a deterministic function of time. Rather than performing steady-state analysis, the Dyna-METRIC approach is to perform a transient analysis at some particular point in time that is chosen by the modeler. The Dyna-METRIC model does not include the possibility to expedite repair. Demand fluctuations are therefore only buffered by holding inventory.

A similar approach is followed by Lau and Song (2008) with two exceptions: They also model the finite repair capacity using queueing approximations and they evaluate the transient behavior of the system at several points of interest rather than only one. For their extensions to Dyna-METRIC, they take heuristic or approximative approaches.

Our work differs from these contributions because demand fluctuations are modeled by a Markov modulated Poisson process. This resembles practice more closely as

the intensity of demand over time behaves as a stochastic process rather than a deterministic function. Additionally, our model deals with these demand fluctuations not only by holding repairable inventory, but also by using the possibility to expedite repair. Our modeling also allows us to evaluate our system exactly and compute tight lower bounds on optimal system performance.

5.2.2 Expediting and repair scheduling policies

The possibility to either expedite repair or prioritize the scheduling of repairs in the repair shop has been considered many times, mostly under the assumption of fixed given turn-around stock levels (Hausman and Scudder, 1982; Scudder, 1986; Scudder and Chua, 1987; Pyke, 1990; Tiemessen and Van Houtum, 2012). In these contributions, the repair shop is modeled by a finite server queue. Given a limited capacity, the question becomes: How should limited repair capacity be allocated to repair jobs of various types, i.e., which repair jobs deserve priority?

As observed by Tiemessen and Van Houtum (2012), even for fixed given turn-around stock levels, computing optimal priority rules, or evaluating a given rule, requires computation times that grow exponentially in the number of different repairable types. Accordingly, most contributions in this area use simulation to study heuristic priority rules. All these authors report that system performance increases substantially by using various priority rules. Hausman and Scudder (1982) and Tiemessen and Van Houtum (2012) both point out that substantial stock reductions should be possible as a result of using an effective priority rule. Under *static* priority rules, the priority of a spare part depends on its type only. Under these relatively simple rules, Sleptchenko et al. (2005) and Adan et al. (2009) have shown numerically that significant reductions in inventory investment are possible compared to simple first come first serve scheduling of repair jobs. More sophisticated priority rules also consider the on-hand inventory and expected future demand in deciding the priority of a part. These *dynamic* priority rules are essentially mechanisms that change the repair lead time of an item based on current on-hand stock and estimated future demand. In this regard, the possibility to schedule repairs can be interpreted as providing lead time flexibility. The expediting policy in our model provides this lead time flexibility, but does not suffer from the tractability issues that dynamic priority queueing models suffer from.

We retain tractability because we assume a rather simple priority rule and refrain from explicitly modeling the queueing behavior that occurs in the repair shop. If the repair shop is external to the company holding inventory, this is a natural modeling choice, but even when the repair shop is internal to the company, this model has merit: In many organizations, the repair shop and inventories are managed

separately. Coordination of repair priorities often happens implicitly through lead time agreements between the inventory manager and the repair shop manager. Our model is a first step in explicitly considering the effect of smart priority rules when deciding on turn-around stocks. We believe it is also useful in practical situations in which a more more sophisticated priority rule is used.

The possibility to expedite the repair of a part without considering queueing effects in the repair shop has been considered previously by Verrijdt et al. (1998), but their policy only depends on the on-hand inventory of a part and considers Poisson demand only. Moinzadeh and Schmidt (1991) study the same policy that we use, but in the context of deterministic lead times and Poisson demand. Song and Zipkin (2009) show that the model of Moinzadeh and Schmidt (1991) can be reinterpreted as a special type of queueing network for which a product-form solution exists. This observation allows them to significantly generalize the model of Moinzadeh and Schmidt (1991), but it does not allow expediting policies that somehow depend on demand fluctuations. The expediting policy we propose in this chapter, does depend on demand fluctuations, and is shown to be optimal under certain conditions described in chapter 4 of this thesis. The merit of this rule is that it captures the essential trade-off involved in dynamically scheduling repair of spare parts, while being sufficiently simple to make the problem of deciding inventory levels tractable.

5.2.3 Decomposition and column generation

Decomposition and column generation is a general technique to deal with optimization problems that have a Lagrangian[1] that can be decomposed. The most straightforward way of dealing with such problems is by manipulating the Lagrange multipliers as suggested by Everett (1963)[2] and later by Fisher (1981). Brooks and Geoffrion (1966) noted that one efficient way of finding the best Lagrange multipliers is via setting up a linear program in which each variable corresponds to a solution for each of the parts

[1]The Lagrangian of a constrained optimization problem

$$\min_{x\in\mathbb{R}^m}\{f(x)|g_i(x)\leq b_i,\quad i=1,...n\}$$

is given by

$$L(x,u)=f(x)+\sum_{i=1}^{n}u_i(g_i(x)-b_i),$$

where x and $u=(u_1,\ldots,u_n)$ are vectors. For all fixed $u\geq 0$, a minimum of $L(x,u)$ is also a lower bound for the original optimization problem. The Lagrangian was called after the Italian mathematician and astronomer Joseph-Louis Lagrange (1736-1813) who developed this technique.

[2]Hugh Everett III (1930-1982) was a quantum physicist who proposed the many worlds interpretation of quantum mechanics. His contribution to operations research, known commonly as the Everett result, was conceived in a hotel room while he was visiting Copenhagen and failed to come to any agreement with Niels Bohr on their differing interpretations of quantum mechanics.

that compose the Lagrangian. The Lagrange multipliers then correspond to shadow prices (or dual variables) of the linear program. Their algorithm is essentially the decomposition and column generation algorithm that we use in this chapter.

In the context of spare parts inventory optimization, decomposition and column generation has been used as early as in the seminal paper of Sherbrooke (1968) to solve the METRIC model, where the Lagrangian is decomposable per spare part type. Essentially, the technique reduces the original optimization problem that encompasses many types of repairables, to repeatedly solving a single-item inventory problem for each repairable. Usage of this technique for spare part inventory optimization problems has found much recent following, e.g. Kranenburg and Van Houtum (2007, 2008); Alvarez et al. (2013a,b). In all these papers (including Sherbrooke (1968)), there is one or more service level constraints that need to be achieved by all parts collectively (rather than individually). After moving these service level constraints to the objective by taking the Lagrangian, the best Lagrange multipliers are found via dual variables in a linear programming relaxation of the problem. (See also Dantzig and Wolfe (1960) and Lübbecke and Desrosiers (2005) for a more general and thorough treatment of this technique.)

We use the same technique to find a lower bound and a feasible solution for our model. Different from all the papers mentioned in the previous paragraph, different repairable items are not only linked because of a collective service level, but also through the expediting load that they have on one or more repair resources. This is a merit of how our model is set up: Our expediting rule mimics the dynamic priorities given to repairs but allows for tractable analysis through the technique of decomposition and column generation.

5.3. Model

In this section, we model our problem and illustrate most modeling steps by continuing the example started in the introduction. We start with some notation and preliminaries in §5.3.1. Then we discuss the control policy we use for each repairable type in §5.3.2. Fluctuating demand models are discussed in §5.3.3. We conclude this section by formally stating our optimization problem in §5.3.4.

5.3.1 Notation and preliminaries

We consider several fleets of assets for which we keep repairable spare parts on stock. We denote the set of fleets by A and the set of repairable items by I. We refer to each element of I as a stock keeping unit (SKU). The set of SKUs used to maintain

fleet $a \in A$ is denoted I_a^A. There is a set of repair resources, C, that are used to repair defective parts. The items that load repair resource $c \in C$ are contained in the set I_c^C. We will assume that $\cup_{c\in C} I_c^C = \cup_{a\in A} I_a^A = I$ and $\cap_{c\in C} I_c^C = \cap_{a\in A} I_a^A = \emptyset$. This assumption is not essential to the analysis, but it does simplify notation and presentation.

Each SKU $i \in I$ faces Markov modulated Poisson demand. This means that demand for SKU i is a Poisson process whose intensity varies with the state of an exogenous Markov process Y_i^t. The Markov process Y_i^t is irreducible and has a finite state space $\Theta_i = \{1, ..., |\Theta_i|\}$ with generator matrix $\mathbf{Q}_i$ whose elements we denote by $q_i(m, n)$. For notational convenience, we define $q_i(m) = -q_i(m, m)$ and $q_i^{\max} = \max_{m\in\Theta_i} q_i(m)$. When $Y_i^t = y$, the intensity of Poisson demand at time t is given by $\lambda_i(y) \geq 0$; $\boldsymbol{\lambda}_i = (\lambda_i(1), ..., \lambda_i(|\Theta|))$, $\lambda_i(y) > 0$ for at least one $y \in \Theta_i$ and $\lambda_i^{\max} = \max_{y\in\Theta_i} \lambda_i(y)$. We denote demand for SKU i in the time interval $(t_1, t_2]$ given $Y_i^{t_1} = y$ as $D_i^y(t_1, t_2)$. Note that $Y_i^{t_1}$ provides information about the distribution of demand in the interval $(t_1, t_2]$, $t_2 > t_1$. We assume that Y_i^t can be observed directly for all $i \in I$ and provides a form of aggregated advance demand information. In the previous chapter, we already discussed the modeling versatility of this demand model. In example 5.3, we will encounter an example of how demand might fluctuate over time. We address how to model such demand and provide examples in §5.3.3.

There exists a regular and an expedited repair option for each SKU $i \in I$. The expedited repair lead time for SKU i is deterministic and denoted by ℓ_i. The expedited repair lead time may represent things such as the transport time and the repair time or a lead time agreed upon with an external company that provides emergency repair service. We also refer to using the expedited repair mode as expediting repair. The regular repair lead time of SKU i consists of the emergency repair lead time ℓ_i, and a random component of length L_i. The random variable L_i has an exponential distribution with mean $1/\mu_i$. L_i models such things as the time that a part waits for resources to become available in the repair shop or the lead time difference between regular and emergency repair lead times as contracted with an external repair shop. The assumption that L_i has an exponential distribution, seems rather restrictive, but numerical evidence in §4.6 suggests that it is not a very strong assumption at all as the performance of the system seems rather insensitive to the exact distribution of L_i for a fixed mean. The inventory manager knows for each repair order of SKU i when L_i has lapsed, and the remaining lead time of an order is ℓ_i.

Of each SKU i, we already own S_i^{LB} parts. The main decision variables are the total number of parts to own for each SKU. This is denoted by S_i for SKU $i \in I$ and is also referred to as the turn-around stock. For each SKU $i \in I$ there is an acquisition price C_i^a for buying additional spare repairables.

Each repair of an SKU $i \in I_c^C$ part, imposes a 'load' of u_i on repair resource $c \in C$.

We use the term 'load' for u_i, but the interpretation of u_i can vary broadly. To illustrate this, consider for example the following two scenarios:

- Repair is performed by an external repair shop and the repair lead time may be shortened in exchange for an increased price for the repair. However, there is a maximum target on the amount of money that can be used for requesting expedited lead times from external parties. In this case, the repair resource c might be this annual target for expedited repairs expenses and u_i is the additional cost of an expedited repair over a regular repair.
- Repairs are conducted by a repair shop within the company. This repair shop can expedite the repair of certain parts upon request, as long as the load imposed on the repair shop by expedited repairs is limited. Manpower is the bottleneck in the repair shop. The load imposed on the repair shop u_i could then be man hours required for the repair of a SKU $i \in I_c^C$ part.

For each repair resource $c \in C$ there is maximum $\mathcal{E}_c^{\max}$ on the load this repair resource is allowed to experience due to expedited repair orders.

Table 5.1 summarizes the notation we have introduced so far as well as notation we will introduce later.

Now we return to our example to put all this notation in some perspective

Example 5.2 Thomas&Co already has a fleet of 200 trains that are used for services with many stops. This fleet is called VILLAGE, while the fleet of 100 trains they are about to buy is called CITY. Now $A = \{\text{CITY}, \text{VILLAGE}\}$. All mechanical repairs are done in an internal repair shop, while the repair of climate and airconditioning units is outsourced to an external company. Therefore, $C = \{\text{OUTSOURCE}, \text{MECHANIC}\}$. Manpower is the bottleneck in the internal repair shop so u_i is measured in man hours if $i \in I^C_{\text{MECHANIC}}$. If $i \in I^C_{\text{OUTSOURCE}}$, then u_i is measured in EUROS. Thomas&Co has gathered all this data as shown in Table 5.2. Note that from Table 5.2 we can also read that $I^C_{\text{MECHANIC}} = \{2, 3, 5, 6\}$, $I^C_{\text{OUTSOURCE}} = \{1, 4\}$, $I^A_{\text{VILLAGE}} = \{1, 2, 3\}$, and $I^A_{\text{CITY}} = \{4, 5, 6\}$. The data not shown in Table 5.2 is that $\ell_i = 2$ and $\mathbb{E}[L_i] = 3$ for all $i \in I$. In the next example, we will consider demand data. ◇

5.3.2 Control policy

Let X_i^t be the number of parts of SKU i that have been sent to regular repair and have not yet completed the exponential phase of their repair at time t. As control policy for each SKU i, we propose to place a replenishment order whenever demand occurs, i.e. we use a $(S_i - 1, S_i)$ replenishment policy. For the expediting policy, we

Table 5.1 Overview of notation

		Sets
I	:	Set of all SKUs.
A	:	Set of all fleets.
C	:	Set of all types of repair shop resources.
I_a^A	:	Set of SKUs used to maintain fleet $a \in A$.
I_c^C	:	Set of SKUs that load repair resource $c \in C$.
Θ_i	:	Set of modulating states of the Markov modulating chain of demand for SKU $i \in I$
		Input Parameters
$\lambda_i(y)$	:	Demand intensity for SKU $i \in I$ when $Y_i(t) = y \in \Theta_i$
$\boldsymbol{\lambda}_i$	:	The vector $(\lambda_i(1), \lambda_i(2), \cdots, \lambda_i(\lvert\Theta_i\rvert))$
$\lambda_i^{\max}$	:	$\max_{y\in\Theta_i} \lambda_i(y)$ for SKU $i \in I$
$\mathbf{Q}_i$	:	Generator matrix of the modulating process $Y_i(t)$ of SKU $i \in I$
$q_i(m,n)$	:	The element of row m column n of $\mathbf{Q}_i$, $i \in I$
$q_i(m)$	:	$-q_i(m,m)$
$q_i^{\max}$	:	$\max_{m\in\Theta_i} q_i(m)$
ℓ_i	:	The (deterministic) expedited repair lead time of SKU $i \in I$
μ_i^{-1}	:	Mean of the additional regular repair lead time, $\mathbb{E}[L_i]$; (the mean regular repair lead time is $\ell_i + \mu_i^{-1}$)
S_i^{LB}	:	Lower bound on the size of the turn-around-stock for SKU $i \in I$
C_i^a	:	Acquisition costs for SKU $i \in I$
u_i	:	Resource load associated with the repair of SKU $i \in I$
$\mathcal{B}_a^{\max}$	:	The maximally allowed mean number of backorders over all SKUs $i \in I_a^A$ for $a \in A$.
$\mathcal{E}_c^{\max}$	:	The maximally allowed mean resource loading resulting from repair expediting over all items $i \in I_c^C$ for expediting resource $c \in C$.
		Decision variables
S_i	:	Size of the turn-around-stock for SKU $i \in I$
$T_i(y)$	:	Expediting threshold for SKU $i \in I$ when $Y_i^t = y \in \Theta_i$
$\mathbf{T}_i$	:	The vector $(T_i(1), T_i(2), \cdots, T_i(\lvert\Theta_i\rvert))$
		Output of model
$X_i^t(S_i, \mathbf{T}_i)$	:	The number of parts of SKU $i \in I$ in regular repair at time t and not arriving to inventory before time $t + \ell_i$ under an expediting policy with thresholds $\mathbf{T}_i$.
$B_i^t(S_i, \mathbf{T}_i)$	:	Random variable that denotes the number of backorders of SKU i at time t under policy $(S_i, \mathbf{T}_i)$;
$D_i^y(t_1, t_2)$	:	Demand for SKU $i \in I$ in the interval $(t_1, t_2]$ given $Y_i^{t_1} = y \in \Theta_i$
L_i	:	Additional regular repair lead time; has exponential distribution with mean μ_i^{-1}
$\mathcal{B}_i(S_i, \mathbf{T}_i)$	:	Expected number of backorders of SKU $i \in I$, $\lim_{t\to\infty} \mathbb{E}[B_i^t(S_i, \mathbf{T}_i)]$
$\mathcal{E}_i(\mathbf{T}_i)$	:	Expected number of repairs of SKU $i \in I$ that are expedited per unit time $\sum_{y\in\Theta_i} \lambda_i(y)\mathbb{P}(X_i(\mathbf{T}_i) \geq T_i(y) \cap Y_i = y)$

Table 5.2 Input data for Thomas&Co

SKU#	Description	C_i^a (kEURO)	Fleet	Repair Resource	u_i	S_i^{LB}
1	Climate unit	30	Village	Outsource	500	2
2	Electro motor	45	Village	Mechanic	16	1
3	Break set	5	Village	Mechanic	4	5
4	Airconditioning unit	10	City	Outsource	500	0
5	Electro motor	30	City	Mechanic	16	0
6	Break set	2	City	Mechanic	4	0

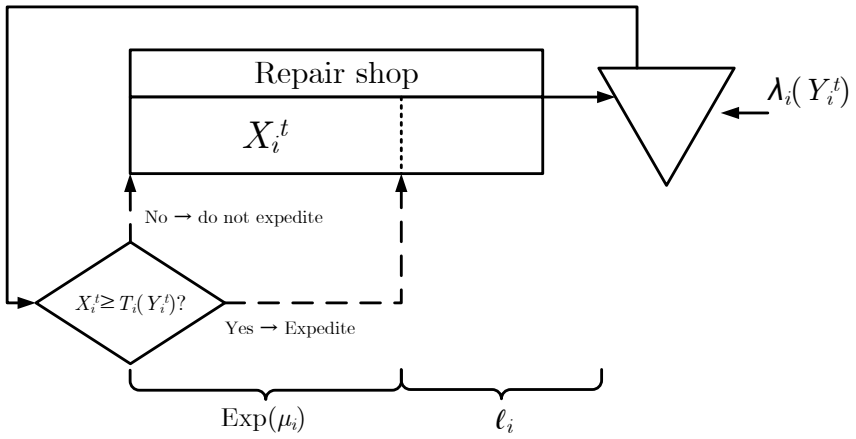

Figure 5.1 A graphical representation of the model for a single item.

propose to expedite whenever X_i^t exceeds some threshold that depends on Y_i^t, i.e. replenishment orders are expedited at time t if $X_i^t \geq T_i(y)$ when $Y_i^t = y$. Thus the control policy for any SKU i can be described by the turn-around stock S_i and a vector $\mathbf{T}_i = (T_i(1), T_i(2), \cdots, T_i(|\Theta_i|))$ containing the expediting thresholds for each modulating state. The stochastic process X_i^t depends on $\mathbf{T}_i$ and so we will write this explicitly: $X_i^t(\mathbf{T}_i)$. Figure 5.1 gives a graphical representation of the control policy for any SKU i. The combined policy is denoted by $(S_i, \mathbf{T}_i)$ and can also be reinterpreted as a state dependent *dual-index policy* as has also been noted in chapter 4. We will use that term for a policy in this chapter.

The state dependent dual-index policy we propose is actually optimal under a linear backordering and expediting cost structure as shown in Theorem 4.1 of chapter 4. Furthermore, the numerical study in §4.6 shows that the performance of a state dependent dual-index policy is rather insensitive to the assumption that L_i has an exponential distribution, i.e. both the performance evaluation error and optimality gap for similar systems where L_i has a different distribution with the same mean are small (within 2.76% and 0.70% respectively over a large test bed).

Under a $(S_i, \mathbf{T}_i)$ policy, $(X_i^t(\mathbf{T}_i), Y_i^t)$ is a Markov process on

$$\mathcal{S} = \left\{ (x, y) \middle| x \in \left\{ 0, \ldots, \max_{k \in \Theta_i} T_i(k) \right\}, \quad y \in \Theta_i \right\}.$$

The Markov process $(X_i^t(\mathbf{T}_i), Y_i^t)$ has three types of transitions:

1. Transitions from (x, y) to $(x+1, y)$ which occur with intensity $\lambda_i(y)$ if $x < T_i(y)$

2. Transitions from (x, y) to $(x-1, y)$ which occur with intensity $x\mu_i$ if $x > 0$
3. Transitions from (x, y) to (x, y') which occur with intensity $q_i(y, y')$ if $y, y' \in \Theta_i$ and $y \neq y'$.

The joint steady state distribution of $(X_i^t(\mathbf{T}_i), Y_i^t)$ can be determined from these transition intensities. When we drop the time superscript t, we refer to the steady state random variables. With the distribution of $(X_i(\mathbf{T}_i), Y_i)$, we can determine the performance of a SKU $i \in I$ in terms of the expected backorders and the expected number of repairs of that are expedited per time unit under policy $(S_i, \mathbf{T}_i)$.

Let $B_i^t(S_i, \mathbf{T}_i)$ denote the number of backorders of SKU $i \in I$ at time t under policy $(S_i, \mathbf{T}_i)$. It satisfies

$$B_i^{t+\ell_i}(S_i, \mathbf{T}_i) = \left(D_i^{Y_i^t}(t, t+\ell_i) - (S_i - X_i^t(\mathbf{T}_i))\right)^+, \tag{5.1}$$

and so the expected number of backorders of SKU $i \in I$ in steady state, $\mathcal{B}_i(S_i, \mathbf{T}_i)$, satisfies:

$$\mathcal{B}_i(S_i, \mathbf{T}_i) = \lim_{t\to\infty} \mathbb{E}\left[B_i^{t+\ell_i}(S_i, \mathbf{T}_i)\right] = \mathbb{E}\left[\left(D_i^{Y_i}(t, t+\ell_i) - S_i + X_i(\mathbf{T}_i)\right)^+\right]. \tag{5.2}$$

Equation (5.2) can be evaluated after noting that $D_i^y(t, t+\ell_i)$ can be computed by numerical inversion of a generating function as explained in §4.A.

Next consider the expected number of repairs that are expedited per time unit of SKU $i \in I$, and denote it by $\mathcal{E}_i(\mathbf{T}_i)$. (Note that $\mathcal{E}_i(\mathbf{T}_i)$ depends on $\mathbf{T}_i$ only, not on S_i.) We have:

$$\mathcal{E}_i(\mathbf{T}_i) = \sum_{y\in\Theta_i} \lambda_i(y)\mathbb{P}(X_i(\mathbf{T}_i) \geq T_i(y) \cap Y_i = y). \tag{5.3}$$

$\mathcal{B}_i(S_i, \mathbf{T}_i)$ and $\mathcal{E}_i(\mathbf{T}_i)$ can be evaluated in many ways. In §5.5, we use value iteration to compute $\mathcal{B}_i(S_i, \mathbf{T}_i)$ and $\mathcal{E}_i(\mathbf{T}_i)$.

5.3.3 Markov Modulated demand models and fitting

Fitting a MMPP demand model to data has not received much attention in the literature. Fitting procedures do exist, but these are geared primarily to applications of queueing models in telecommunication systems (e.g. Heffes and Lucantoni, 1986; Meier-Hellstern, 1987; Yoshihara et al., 2001; Nelson and Gerhardt, 2010). Using Markov modulated demand in the context of inventory problems has been advocated by Song and Zipkin (1993) and Zipkin (2000). These authors emphasize that this demand model allows for great modeling flexibility and indicate that this demand

model can accommodate such divers phenomena as weather conditions and economic conditions. However, practical algorithms to fit MMPP demand models to data have not been provided in the literature. In this section, we provide two fitting techniques. The first fitting procedure in §5.3.3.1 is specific for the maintenance context of this thesis. The second fitting procedure in §5.3.3.2 is a moment fitting procedure, that we believe can also be useful outside of the setting considered in this chapter.

5.3.3.1 Fitting based on maintenance strategy and installed base

The fitting procedure we describe is best understood by first considering an example.

Example 5.3 For the SKUs in Table 5.2, Maintenance engineers at Thomas&Co are asked to assess what the demand will behave like over the next 30-40 years. From past experience, they know that break sets need to be replaced on each train approximately every year and so they expect a relatively steady demand of $200/50 = 4$ for SKU 3 and $100/50 = 2$ parts per week for SKU 6. (We work with a year of 50 weeks.) An airconditioning unit (SKU 4) is estimated to fail due to random causes about once every 5 years. Over the entire fleet, this means that demand due to failure maintenance will be about $\frac{1}{5}100/50 = 0.4$ parts per week. Additionally, the maintenance engineers expect that the airconditioning units of the entire CITY fleet will need to be overhauled roughly every 4 years. They warn that this will lead to peaks in demand during overhaul periods. How high this peak will be, depends on the length of the overhaul period. Currently, revision periods are planned to last a year. For SKU 5, the CITY electro motor, random failures occur around once every 10 years so they expect a relatively steady demand of $\frac{1}{10}100/50 = 0.2$ per week. Electro motors require overhaul every 7 or so years, so here too, maintenance engineers insist that inventory will be needed to deal with peak demand during overhaul periods. Similar estimates are also available for SKUs 1 and 2: SKU 1 and 2 fail due to random causes once every 4 and 8 years respectively and need to be replaced and overhauled every 4 and 6 years respectively. ◇

Example 5.3 illustrates how an understanding of maintenance can improve the understanding of how demand for certain repairables fluctuates. This understanding can then be modeled as the modulating chain for demand. Suppose that demand for repairables behaves as described in Example 5.3: Demand is relatively steady over some period, until demand peaks because of a revision period in which parts are overhauled preventively. Then a simple MMPP that models demand is the following. Let N_a denote the number of equipment in fleet $a \in A$ and consider an SKU $i \in I_a^A$. Let λ_i^{ran} denote the intensity with which any piece of equipment in the fleet fails randomly (i.e. not due to wear out). Wear out failures do not occur because all

repairables in the fleet are overhauled during revision periods. The time between revision periods is a random variable M_i for SKU i. (M_i is not deterministic because the time between revision periods is decided upon based on the condition of the fleet.) Once the revision period starts, it lasts R_i time units and all repairables in the fleet are expected to be replaced and revised during this period. R_i is also a random variable. If we approximate M_i and R_i by exponential random variables a MMPP demand model is given by:

$$\mathbf{Q}_i = \begin{pmatrix} -\mathbb{E}[M_i]^{-1} & \mathbb{E}[M_i]^{-1} \\ \mathbb{E}[R_i]^{-1} & -\mathbb{E}[R_i]^{-1} \end{pmatrix}, \quad \boldsymbol{\lambda}_i^{\mathrm{T}} = \begin{pmatrix} \lambda_i^{\mathrm{ran}} N_a \\ \lambda_i^{\mathrm{ran}} N_a + N_a/\mathbb{E}[R_i] \end{pmatrix}. \tag{5.4}$$

Rather than using the exponential distribution for R_i and M_i, it is possible to use any phase type distribution if appropriate. The restriction of modeling R_i and M_i by phase type distributions is rather weak because phase type distributions are dense in the class of all non-negative distributions (Schassberger, 1973). If we choose to model M_i by an Erlang-2 distribution, we obtain:

$$\mathbf{Q}_i = \begin{pmatrix} -(\frac{1}{2}\mathbb{E}[M_i])^{-1} & (\frac{1}{2}\mathbb{E}[M_i])^{-1} & 0 \\ 0 & -(\frac{1}{2}\mathbb{E}[M_i])^{-1} & (\frac{1}{2}\mathbb{E}[M_i])^{-1} \\ \mathbb{E}[R_i]^{-1} & 0 & -\mathbb{E}[R_i]^{-1} \end{pmatrix}, \quad \boldsymbol{\lambda}_i^{\mathrm{T}} = \begin{pmatrix} \lambda_i^{\mathrm{ran}} N_a \\ \lambda_i^{\mathrm{ran}} N_a \\ \lambda_i^{\mathrm{ran}} N_a + \frac{N_a}{\mathbb{E}[R_i]} \end{pmatrix}. \tag{5.5}$$

Example 5.4 Thomas&Co decide to use (5.4) to model their demand. This yields (time units are weeks):

$$\mathbf{Q}_1 = \begin{pmatrix} -\frac{1}{200} & \frac{1}{200} \\ \frac{1}{50} & -\frac{1}{50} \end{pmatrix}, \qquad \mathbf{Q}_3 = 0, \qquad \mathbf{Q}_5 = \begin{pmatrix} -\frac{1}{350} & \frac{1}{350} \\ \frac{1}{50} & -\frac{1}{50} \end{pmatrix}$$
$$\mathbf{Q}_2 = \begin{pmatrix} -\frac{1}{400} & \frac{1}{400} \\ \frac{1}{50} & -\frac{1}{50} \end{pmatrix}, \qquad \mathbf{Q}_4 = \mathbf{Q}_1, \qquad \mathbf{Q}_6 = 0,$$

and

$$\boldsymbol{\lambda}_1^{\mathrm{T}} = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \boldsymbol{\lambda}_2^{\mathrm{T}} = \begin{pmatrix} \frac{1}{2} \\ \frac{9}{2} \end{pmatrix}, \boldsymbol{\lambda}_3^{\mathrm{T}} = 4, \boldsymbol{\lambda}_4^{\mathrm{T}} = \begin{pmatrix} \frac{2}{5} \\ \frac{12}{5} \end{pmatrix}, \boldsymbol{\lambda}_4^{\mathrm{T}} = \begin{pmatrix} \frac{1}{5} \\ \frac{11}{5} \end{pmatrix}, \boldsymbol{\lambda}_6^{\mathrm{T}} = 2.$$

◇

5.3.3.2 Fitting based on moments of demand over expected lead time

One of the drawbacks of the stationary Poisson demand model is that it has only one parameter and so fixing the mean demand per period, also fixes the variance of demand per period. The MMPP allows for fitting arbitrary moments of demand for

any finite time span provided that the variation coefficient (variance of demand in that time span divided by the mean demand in that time span) is greater than 1.

Suppose we are given the mean $\mathbb{E}[D]$ and variance $\mathbf{Var}[D]$ of demand over some finite time span. We scale time such that this time span is exactly one time unit. The following proposition provides a two-state MMPP that fits these moments.

Proposition 5.1 *Let X be a random variable with mean μ and standard deviation σ that satisfy $\sigma^2/\mu > 1$. The number of counts during one time unit of a MMPP in steady state with parameters*

$$\mathbf{Q} = \begin{pmatrix} -\beta & \beta \\ \alpha\beta & -\alpha\beta \end{pmatrix}, \qquad \boldsymbol{\lambda} = (0, \lambda), \tag{5.6}$$

matches the first two moments of X if α is fixed to verify

$$\alpha \geq \kappa \frac{\sigma^2 - \mu}{\mu^2}, \tag{5.7}$$

for some $\kappa \geq 2$, λ is fixed as

$$\lambda = (1 + \alpha)\mu, \tag{5.8}$$

and β is fixed as the unique and attractive solution to the fixed point equation

$$\beta = \frac{\mu\sqrt{2\alpha e^{-(\alpha+1)\beta}(\sigma^2 - \mu) + 2\alpha(\mu - \sigma^2) + \alpha^2\mu^2} + \alpha\mu^2}{(\alpha + 1)(\sigma^2 - \mu)}. \tag{5.9}$$

Appendix 5.A provides the proof of Proposition 5.1 as well as several figures of the fit that this procedure provides.

5.3.4 Optimization problem

The objective of the manager is to minimize the investment he is about to make in buying repairable spare parts. The constraints are to keep the total expected backorders for each fleet $a \in A$ below $\mathcal{B}_a^{\max}$ and to keep the total expected resource loading due to expedited repair orders below $\mathcal{E}_c^{\max}$ for each repair resource $c \in C$. A backorder for a part renders some equipment down. If an expedited repair mode is available for SKU $i \in I$, it is unacceptable that any particular backorder for SKU $i \in I$ lasts longer than ℓ_i. To ensure this never happens, it suffices to ensure that $T_i(y) \leq S_i$ for each $i \in I$ and $y \in \Theta_i$. Combining all this leads to the following formal

statement of our optimization problem which we call P:

$$
\begin{aligned}
(P) \quad \min_{\{S_i, \mathbf{T}_i \mid i \in I\}} \quad & \sum_{i \in I} C_i^a \left(S_i - S_i^{LB}\right) && && (5.10)\\
\text{subject to} \quad & \sum_{i \in I_a^A} \mathcal{B}_i(S_i, \mathbf{T}_i) \leq \mathcal{B}_a^{\max} && \forall a \in A && (5.11)\\
& \sum_{i \in I_c^C} u_i \mathcal{E}_i(\mathbf{T}_i) \leq \mathcal{E}_c^{\max} && \forall c \in C && (5.12)\\
& S_i^{LB} \leq S_i && \forall i \in I && (5.13)\\
& T_i(y) \leq S_i && \forall i \in I, \forall y \in \Theta_i && (5.14)\\
& S_i, T_i(y) \in \mathbb{N}_0 && \forall i \in I, \forall y \in \Theta_i. && (5.15)
\end{aligned}
$$

We denote the optimal costs to problem (P) by C_P. In the next section, we construct a feasible solution with cost C_P^{UB} for problem (P) as well as a lower bound, C_P^{LB}, on the optimal cost of problem (P). Section 5.4 can be skipped without loss of continuity.

Example 5.5 Thomas&Co would like to adhere to the goals of having $\mathcal{B}_{\text{VILLAGE}}^{\max} = 1$ and $\mathcal{B}_{\text{VILLAGE}}^{\max} = 0.5$. For expediting the repair of climate and airconditioning units (OUTSOURCE repair resource) there is a weekly budget of 200 EUROS, $\mathcal{E}_{\text{OUTSOURCE}}^{\max} = 200$. (Note that the 'loads' for each SKU $i \in I$ are provided in Table 5.2 as discussed in Example 5.2.) For expediting the repair for the internal repair shop that handles mechanical repairs, the agreement with the repair shop manager is to keep requests for expedited repair orders below the nominal load of 20 man hours per week on average, $\mathcal{E}_{\text{MECHANIC}}^{\max} = 20$. ◇

5.4. Analysis

The analysis will proceed by giving an algorithm to construct a lower bound for problem (P) in §5.4.1. In 5.4.2, we show how to find a good feasible solution for problem (P) based on the lower bound constructed in §5.4.1.

5.4.1 Constructing lower bounds with column generation

To obtain a lower bound for problem (P), we first reformulate it to an integer linear program and then relax the integrality constraints. We refer to this problem as the master problem (MP). To this end, we introduce the set K_i of all dual-index policies k for item i that respect constraints (5.13)-(5.15) of problem (P). Policy $k \in K_i$ has base-stock level and expediting thresholds $(S_i^k, \mathbf{T}_i^k)$. We also introduce the decision

variable $x_i^k \in \{0,1\}$ that indicates whether policy k is chosen for item i. If we relax the integrality constraint on x_i^k, we obtain the master problem:

$$
\begin{aligned}
(MP) \quad \min_{\{x_i^k | i\in I, k\in K_i\}} \quad & \sum_{i\in I} C_i^a \left(S_i^k - S_i^{LB}\right) x_i^k && && (5.16)\\
\text{subject to} \quad & \sum_{i\in I_a^A}\sum_{k\in K_i} \mathcal{B}_i(S_i^k, \mathbf{T}_i^k) x_i^k \le \mathcal{B}_a^{\max} && \forall a \in A && (5.17)\\
& \sum_{i\in I_c^C}\sum_{k\in K_i} u_i \mathcal{E}_i(\mathbf{T}_i^k) x_i^k \le \mathcal{E}_c^{\max} && \forall c \in C && (5.18)\\
& \sum_{k\in K_i} x_i^k = 1 && \forall i \in I && (5.19)\\
& x_i^k \ge 0 && \forall i \in I, \forall k \in K_i.
\end{aligned}
$$

Since K_i is an infinite set, (MP) is an infinite dimensional linear program. The way to solve (MP), is to introduce a restricted master problem (RMP) in which we replace K_i with a finite subset K_i^{res} and solve (RMP) to optimality. Then we consider whether we can improve the solution to (RMP) by adding policies $k \in K_i \setminus K_i^{\text{res}}$ to K_i^{res}. To see if such policies exist for SKU i, we need to solve a sub-problem. (This sub-problem is also called the column generation problem or pricing problem.) We let p_a denote the dual variable of (RMP) corresponding with fleet $a \in A$ for constraint (5.17), ρ_c denote the dual variable of (RMP) corresponding with repair resource $c \in C$ for constraint (5.18) and v_i denote the dual variable of (RMP) corresponding with SKU i for constraint (5.19). If $i \in I_a^A \cap I_c^C$, then the sub-problem for SKU i is given by:

$$
\begin{aligned}
(SUB(i)) \quad \min_{\{(S_i, \mathbf{T}_i)\}} \quad & C_i^a \left(S_i - S_i^{LB}\right) - p_a \mathcal{B}_i(S_i, \mathbf{T}_i) - \rho_c u_i \mathcal{E}_i(\mathbf{T}_i) - v_i\\
\text{subject to} \quad & S_i^{LB} \le S_i\\
& T_i(y) \le S_i && \forall y \in \Theta_i && (5.20)\\
& S_i, T_i(y) \in \mathbb{N}_0 && \forall y \in \Theta_i. && (5.21)
\end{aligned}
$$

If a feasible solution to $(SUB(i))$ exists with a negative objective value, then the objective of (RMP) can be improved by adding this solution to K_i^{res} and solving (RMP) with this larger set K_i^{res}. An optimal solution to (RMP) is also an optimal solution for (MP) if the optimal objective of $(SUB(i))$ is non-negative for each $i \in I$. Since (MP) is a relaxation of (P), we have also found a lower bound for problem (P) that we denote by C_P^{LB}.

Note that all policies that yield a negative objective for $(SUB(i))$ can improve the solution of (RMP), so we do not need to solve $(SUB(i))$ to optimality each time we obtain new dual variables from the restricted master problem. We do need to solve

$(SUB(i))$ to optimality to verify that an optimal solution to (RMP) is also optimal for (MP). The next section treats heuristic and exact methods to solve $(SUB(i))$.

5.4.1.1 Solving the sub-problem

The optimization problem $(SUB(i))$ is almost identical to the single-item problem discussed in chapter 4. The main differences are that:

- $(SUB(i))$ assumes a state dependent dual-index form for the control policy for each item
- The expediting thresholds in $(SUB(i))$ are restricted to be below S_i rather than any number in $\mathbb{N}_0 \cap \{\infty\}$.

However, the methods from chapter 4 can be applied almost immediately by observing that the form of the policy we assume is actually optimal as shown in Theorem 4.1 and that constraint (5.20) can be accommodated by setting the constant M in chapter 4 equal to S_i. The exact and heuristic methods in §4.4 and §4.5 can easily be adapted to solve $(SUB(i))$ by restricting the search over S_i to be above S_i^{LB}. The exact method in §4.4 is also exact for $(SUB(i))$ because the bounds on an optimal S_i given in Proposition 4.2 are obtained via lower and upper bounds that are also lower and upper bounds on the optimal objective of $(SUB(i))$.

5.4.2 Constructing a good feasible solution

Several methods have been suggested to find a good feasible solution based on a lower bound of the type constructed in the previous section. Kranenburg and Van Houtum (2007) and Kranenburg and Van Houtum (2008) suggest rounding the fractional solution obtained from solving (MP) and then performing a local search to find a good feasible solution. More recently, Alvarez et al. (2013a) and Alvarez et al. (2013b) suggest solving the final version of (RMP) after all columns have been generated as an integer linear program. Because the found very good results compared to local search algorithms, we also take that approach. To speed up the solution process we use the feasibility pump heuristic (Fischetti et al., 2005) and stop the solution of the integer linear program as soon as a feasible solution with optimality gap[3] of less than 0.5% is found or 1 minute has elapsed (whichever occurs first). This results in a feasible solution to (P) that is also an upper bound. We denote the cost of this solution by C_P^{UB}.

[3]Observe that this optimality gap is with respect to the integer linear programming formulation with a finite number of columns, *not* with respect to the original optimization problem.

Alvarez et al. (2013a) and Alvarez et al. (2013b) report that this approach is computationally feasible with a commercial solver such as CPLEX. Our approach works well with the GLPK open source solver, even though the performance of this solver is consistently lagging in benchmarks[4].

Example 5.6 For the instance of Thomas&Co, we find a lower bound on the optimal cost of $C_P^{LB} = 851.58$ kEURO. (Note that since all prices of parts are integer multiples of 1000 EURO, 852 kEURO is also a lower bound on the optimal costs of acquiring new repairable parts.) We also found a feasible solution with cost $C_P^{UB} = 892$ kEURO. This solution is shown in Table 5.3. The solution in Table 5.3 is further characterized

Table 5.3 Feasible solution for the Thomas&Co problem (P)

SKU#	S_i	$T_i(1)$	$T_i(2)$
1	19	19	11
2	4	2	0
3	12	4	-
4	12	12	12
5	2	1	0
6	16	9	-

by $\sum_{i \in I^A_{\text{VILLAGE}}} \mathcal{B}_i(S_i, \mathbf{T}_i) = 0.940$, $\sum_{i \in I^A_{\text{CITY}}} \mathcal{B}_i(S_i, \mathbf{T}_i) = 0.485$, $\sum_{i \in I^C_{\text{OUTSOURCE}}} \mathcal{E}_i(\mathbf{T}_i) = 176.231$, and $\sum_{i \in I^C_{\text{MECHANIC}}} \mathcal{E}_i(\mathbf{T}_i) = 19.996$. The optimality gap $(C_P^{UB} - C_P^{LB})/C_P^{LB} \cdot 100\% = 4.7\%$.

5.5. Computational results

We discuss the questions we would like to answer, and the test bed we use in §5.5.1. We present and discuss the numerical results in §5.5.2.

5.5.1 Objectives and test bed

The objectives of this numerical study are to:

1. Determine whether the algorithm to find a feasible solution to (P) is effective, i.e., determine whether it finds solutions that are close to optimal;
2. Determine whether the algorithm to find a feasible solution to (P) is efficient, i.e., determine whether it finds a feasible solution within reasonable time;

[4]See for example the MIPLIB2010 (Koch et al., 2011) benchmark accessible via the benchmark site of Hans Mittelmann: `http://plato.asu.edu/bench.html`

3. Determine by how much stock investment can be reduced because of the possibility to expedite repair of parts.

To answer these question, we set up a large test bed of instances. The order of magnitude of problem parameters for our test instances are based on observations made at NedTrain. We introduce the notation $U(a, b)$ for a uniform random variable on the interval (a, b). An overview of how instances in the test bed are generated is shown in Table 5.4. The total number of instances in the test bed is $3^5 2^3 = 1944$. For each combination of parameters 1,2,3,4,5,9, and 10 in Table 5.4, we generate two instances randomly as follows:

- For each SKU $i \in I$, we generate a Markov modulated Poisson demand process with $\mathbf{Q}$ generated as shown under 7 in Table 5.4, and $\boldsymbol{\lambda}$ generated by one of the two option shown under 8 in Table 5.4. (This is why two instances are generated.);
- Each SKU $i \in I$ is assigned uniformly at random to a repair resource set I_c^C for $c = 1, \ldots, |C|$;
- For each SKU $i \in I$, we generate an acquisition price from $U(100, 1000)$;
- For each SKU $i \in I$, set $u_i = 1$;
- The values $\mathcal{B}_a^{\max}$ and $\mathcal{E}_c^{\max}$ are set as fractions ν and ξ of the the total expected demand per time unit for the fleet and repair resource respectively as shown under 9 and 10 in Table 5.4.

To assess the value of expediting, we create a 'benchmarking' instance for each 'orginal' instance of (P) that we generate. This benchmarking instance is created to be identical to the original instance except that the mean repair lead time of the benchmark instance is less than or equal to te mean repair lead time of the original instance, but such that it is not possible to differentiate repair lead times through expediting. This is achieved as follows. We raise $\mathcal{E}_c^{\max}$ for each $c \in C$ of the original instance such that it is feasible (and optimal) to expedite all repairs. We change the expedited lead time to the shortest possible mean repair lead time possible in the original instance which is $\xi\ell_i + (1 - \xi)(\ell_i + \mathbb{E}[L_i])$. (Note that this procedure works because $u_i = 1$ for all $i \in I$.)

Now for each generated original instance we compute a feasible solution with cost C_P^{UB} as described in §5.4.2 and compare it to the lower bound C_P^{LB} that is obtained via the method described in §5.4.1:

$$\%GAP = \frac{C_P^{UB} - C_P^{LB}}{C_P^{LB}} \cdot 100\%. \tag{5.22}$$

Table 5.4 Parameters for test bed instances

	Parameter	Values
1	Number of fleets $\|A\|$	1,2,4
2	Number of repair resources $\|C\|$	1,2,4
3	Number of SKUs per fleet $\|I_a^A\|$	20,50,100
4	Mean of additional regular repair lead time, $\mathbb{E}[L_i]$	2,4
5	Expedited repair lead time, ℓ_i	1,2
6	Acquistion cost for SKU $i \in I$, C_i^a	$U(100, 1000)$
7	Modulating chain generator for SKU $i \in I$, $\mathbf{Q}_i$	$\begin{pmatrix} -q_1 & q_1 \\ q_2 & -q_2 \end{pmatrix}$ with $q_1 = [U(200, 400)]^{-1}$, $q_2 = [U(5, 50)]^{-1}$
8	Demand intensity vector for SKU $i \in I$, $\boldsymbol{\lambda}_i$	$\begin{pmatrix} U(0.01, 0.1) \\ U(0.5, 1.5) \end{pmatrix}$, $\begin{pmatrix} U(0.01, 0.5) \\ U(1, 2) \end{pmatrix}$,
9	Upper bound on backorders for fleet $a \in A$, $\mathcal{B}_a^{\max}$	$\nu \sum_{i \in I_a^A} \sum_{y \in \Theta_i} \mathbb{P}(Y_i = y)\lambda_i(y)$ for $\nu = 0.05, 0.02, 0.01$
10	Upper bound on expediting load for resource $c \in C$, $\mathcal{E}_c^{\max}$	$\xi \sum_{i \in I_c^C} \sum_{y \in \Theta_i} \mathbb{P}(Y_i = y)\lambda_i(y)$ for $\xi = 0.2, 0.1, 0.05$

Next we investigate the relative difference with the benchmark instance. We denote a lower bound on the optimal objective of the benchmark instance by C_{BENCH}^{LB}. We compare C_P^{UB} with C_{BENCH}^{LB}:

$$\%VAL = \frac{C_{BENCH}^{LB} - C_P^{UB}}{C_{BENCH}^{LB}} \cdot 100\%. \tag{5.23}$$

The algorithms described in §§5.4.1-5.4.2 were programmed as a single threaded application in C with GLPK as the solver of both linear and integer linear programs. All computations were carried out on a PC running Windows (32 bit) with Intel Core Duo 2.33 GHz CPU and 4 GB of RAM.

5.5.2 Results

Table 5.5 shows the results of the computational experiment. For each of the parameters in Table 5.4 that has several settings, we computed the mean and maximum $\%GAP$ and $\%VAL$ as well as the mean and maximum computation time in seconds for each of the settings. We will now discuss objective 1-3 as stated in the previous subsection.

The average optimality gap of our feasible solution is very small at 0.67% but optimality gaps of up to 6.76% do occur. The optimality gap seems to increase with the number of fleets and repair resources. This is not surprising, because (MP) has $|I|+|A|+|C|$ constraints and the same number of basic variables in an optimal solution.

Table 5.5 Summary of computational results

		%GAP		%VAL		CPU time (s)	
Parameter	Values	avg	max	avg	max	avg	max
Number of fleets, $\|A\|$	1	0.39	3.00	25.0	48.1	31	154
	2	0.56	6.76	25.0	49.1	76	313
	4	1.05	5.49	24.8	48.2	152	522
Number of repair resources, $\|C\|$	1	0.40	4.54	25.1	49.1	76	473
	2	0.55	3.57	25.1	48.2	85	511
	4	1.04	6.76	24.7	48.1	98	522
Number of SKUs per fleet, $\|I_a^A\|$	20	0.64	5.49	24.7	49.1	38	157
	50	0.68	4.71	25.0	47.9	80	282
	100	0.68	6.76	25.1	47.6	141	522
Fraction of total demand per time unit that may be backordered, ν	0.05	0.72	6.76	24.9	48.2	78	462
	0.02	0.68	5.49	24.8	48.0	88	502
	0.01	0.61	5.26	25.1	49.1	93	522
Fraction of total demand per time unit that may be expedited, ξ	0.2	0.88	6.76	32.2	49.1	86	452
	0.1	0.62	4.71	24.3	38.8	88	502
	0.05	0.50	3.72	18.3	31.4	85	522
Expedited repair lead time, ℓ_i	1	0.70	6.76	28.5	49.1	59	305
	2	0.63	5.26	21.4	42.7	113	522
Random demand	$\begin{pmatrix} U(0.01,0.5) \\ U(1,2) \end{pmatrix}$	0.72	5.26	28.2	49.1	111	522
intensity vector	$\begin{pmatrix} U(0.01,0.1) \\ U(0.5,1.5) \end{pmatrix}$	0.61	6.76	21.7	41.4	62	281
Additional regular repair lead time, $\mathbb{E}[L_i]$	2	0.69	4.74	20.9	39.4	82	502
	4	0.64	6.76	29.0	49.1	91	522
Total		0.67	6.76	24.9	49.1	86	522

Because of constraint (5.19), there is a basic variable for each $i \in I$. Therefore, there will be at most $|A| + |C|$ SKUs for which the optimal solution to (MP) is fractional. This explains why the optimality gap increases with both $|A|$ and $|C|$. Somewhat surprisingly, the optimality gap does not seem to decrease significantly with $|I_a^A|$. This is different form other multi-item spare parts problem where the optimality gap typically does decrease with the number of SKUs considered, (e.g Kranenburg and Van Houtum, 2007, 2008; Alvarez et al., 2013a,b). This can be explained by the fact that we put a time limit of 1 minute on the integer linear programming solver.

The computation times of finding a feasible solution are 86 seconds on average and at most 522 seconds, which is quite acceptable given the size of the problems. It is also convenient that the computation time seems to scale linearly in the number of SKUs. Over 95% of the computation time for solving (MP) to optimality is spent in solving

($SUB(i)$). This task could also be parallelized on modern multi-core processors so that the computation time can be further reduced by a factor equal to the number of cores on a processor.

The value of using expediting to influence repair lead times of repairables is quite valuable with an average benefit of 24.9% and even benefits of up to 49.1%. As was to be expected, the benefits increase with the fraction of total demand that can be expedited and with the expedited lead time. But even the opportunity to expedite 5% of demand leads to average savings of as much as 18.3% compared to static lead times.

5.6. Conclusion

This chapter presented an efficient and effective algorithm to determine near optimal turn-around stock levels for a large group of repairable items that are used in the maintenance of several fleets of equipment. The use of expediting to influence the repair lead time of repairables was shown to be quite effective in reducing the stock investment needed to meet service levels for several fleets of equipment.

5.A. Proof of Proposition 5.1

We start with some preliminaries. Consider a two state MMPP with generator $\mathbf{R}$ and intensity vector $\boldsymbol{\nu}$ given by

$$\mathbf{R} = \begin{pmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{pmatrix}, \qquad \boldsymbol{\nu} = (\nu_1, \nu_2).$$

We let N_t denote the number of arrivals this MMPP generates in an interval of length t when it is in steady state. From Heffes and Lucantoni (1986), we have that

$$\mathbb{E}[N_t] = \frac{\nu_1 r_2 + \nu_2 r_1}{r_1 + r_2} \tag{5.24}$$

and

$$\mathbf{Var}[N_t] = \mathbb{E}[N_t] + 2At - \frac{2A}{r_1 + r_2}\left(1 - e^{-(r_1+r_2)t}\right) \tag{5.25}$$

with

$$A = \frac{r_1 r_2 (\nu_1 - \nu_2)^2}{(r_1 + r_2)^3}.$$

Now we start the proof of Proposition 5.1.

Proof: Let N denote the number of arrivals during one time unit in steady state in the MMPP in the proposition. Using (5.24), we find that

$$\mathbb{E}[N] = \frac{\lambda}{\alpha + 1}, \tag{5.26}$$

and equating this with μ and solving for λ yields

$$\lambda = (\alpha + 1)\mu. \tag{5.27}$$

Substituting (5.27) with $\mathbb{E}[N] = \mu$ into (5.25) yields

$$\mathbf{Var}[N] = \mu + \frac{2\alpha\mu^2}{(\alpha+1)\beta} - \frac{2\alpha\mu^2}{(\alpha+1)^2\beta^2}\left(1 - e^{-(\alpha+1)\beta}\right). \tag{5.28}$$

Equating (5.28) with σ^2, and rearranging we obtain

$$(\sigma^2 - \mu)(\alpha+1)\beta^2 - 2\alpha\mu^2(\alpha+1)\beta + 2\alpha\mu^2 = 2\alpha\mu^2 e^{-(\alpha+1)\beta}. \tag{5.29}$$

Applying the quadratic root formula to (5.29) and simplifying, we find that if there is a $\beta > 0$ that satisfies

$$\beta = \frac{\mu\sqrt{2\alpha e^{-(\alpha+1)\beta}(\sigma^2 - \mu) + 2\alpha(\mu - \sigma^2) + \alpha^2\mu^2} + \alpha\mu^2}{(\alpha+1)(\sigma^2 - \mu)}, \tag{5.30}$$

we have a fit. Now we show that such a unique $\beta^* > 0$ does exist provided

$$\alpha \geq \kappa \frac{\sigma^2 - \mu}{\mu^2}, \quad \text{and} \quad \frac{\sigma^2}{\mu} > 1, \tag{5.31}$$

for some $\kappa \geq 2$.

For convenience define $f : \mathbb{R}_+ \to \mathbb{R}$ as

$$f(\beta) = \frac{\mu\sqrt{2\alpha e^{-(\alpha+1)\beta}(\sigma^2 - \mu) + 2\alpha(\mu - \sigma^2) + \alpha^2\mu^2} + \alpha\mu^2}{(\alpha + 1)(\sigma^2 - \mu)} \tag{5.32}$$

where $\mathbb{R}_+ = [0, \infty)$ and let α, σ^2 and μ satisfy (5.31). To show that there is a unique $\beta^* > 0$ that solves (5.30), it suffices to show that $f(0) > 0$ and that $f'(\beta) < 0$ for all $\beta \in \mathbb{R}_+$. That $f(0) > 0$ can be verified directly and for $f'(\beta)$ we have

$$f'(\beta) = -\frac{\alpha\mu e^{-(\alpha+1)\beta}}{\sqrt{\alpha^2\mu^2 - 2\alpha\left(1 - e^{-(\alpha+1)\beta}\right)(\sigma^2 - \mu)}} < 0. \tag{5.33}$$

The strict inequality holds because (5.31) holds. Next we observe that $f'(\beta) > -1$ for all $\beta > 0$ and in particular for β^*, because $f''(\beta) > 0$ for all $\beta > 0$:

$$\begin{aligned} f''(\beta) = {} & \frac{\alpha\mu e^{-\frac{3}{2}(\alpha+1)\beta}\left\{2\alpha(\alpha+1)e^{(\alpha+1)\beta}(\mu - \sigma^2) + (\alpha+1)\alpha^2\mu^2 e^{(\alpha+1)\beta}\right\}}{2\left\{\alpha^2\mu^2 e^{(\alpha+1)\beta} + 2\alpha e^{(\alpha+1)\beta}(\mu - \sigma^2) + 2\alpha(\sigma^2 - \mu)\right\}^{\frac{3}{2}}} \\ & + \frac{(\alpha+1)\alpha\mu e^{(\alpha+1)\beta}}{\sqrt{\alpha^2\mu^2 e^{(\alpha+1)\beta} + 2\alpha e^{(\alpha+1)\beta}(\mu - \sigma^2) + 2\alpha(\sigma^2 - \mu)}} > 0. \end{aligned} \tag{5.34}$$

The strict inequality again holds because (5.31) holds. Since $f'(0) = -1$, and $f'(\beta) < 0$ and $f''(\beta) > 0$ for all $\beta > 0$, we conclude that $|f'(\beta)| < 1$ for all $\beta > 0$ and in particular for β^*. This implies that β^* is an attractive fixed point of f. □

The fit provided in Proposition 5.1 is parameterized by $\kappa \geq 2$. To gain some intuition on the fit provided and the role of the parameter κ, we provide some examples of the distribution of N_1 that this fit generates in Figures 5.2 and 5.3.

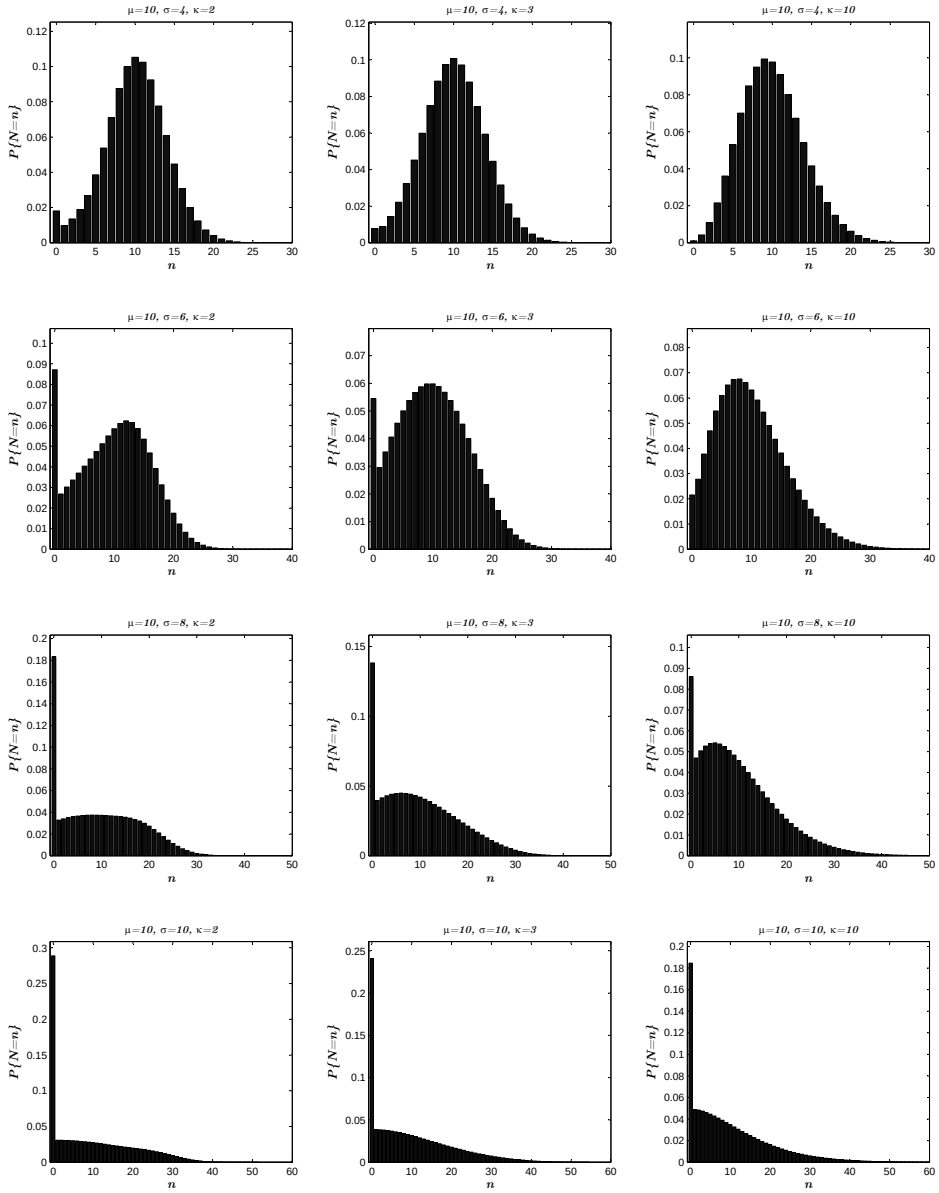

Figure 5.2 Fitted distributions generated by the procedure in Proposition 5.1. Standard deviation and the fitting parameter κ are varied as shown in the plots.

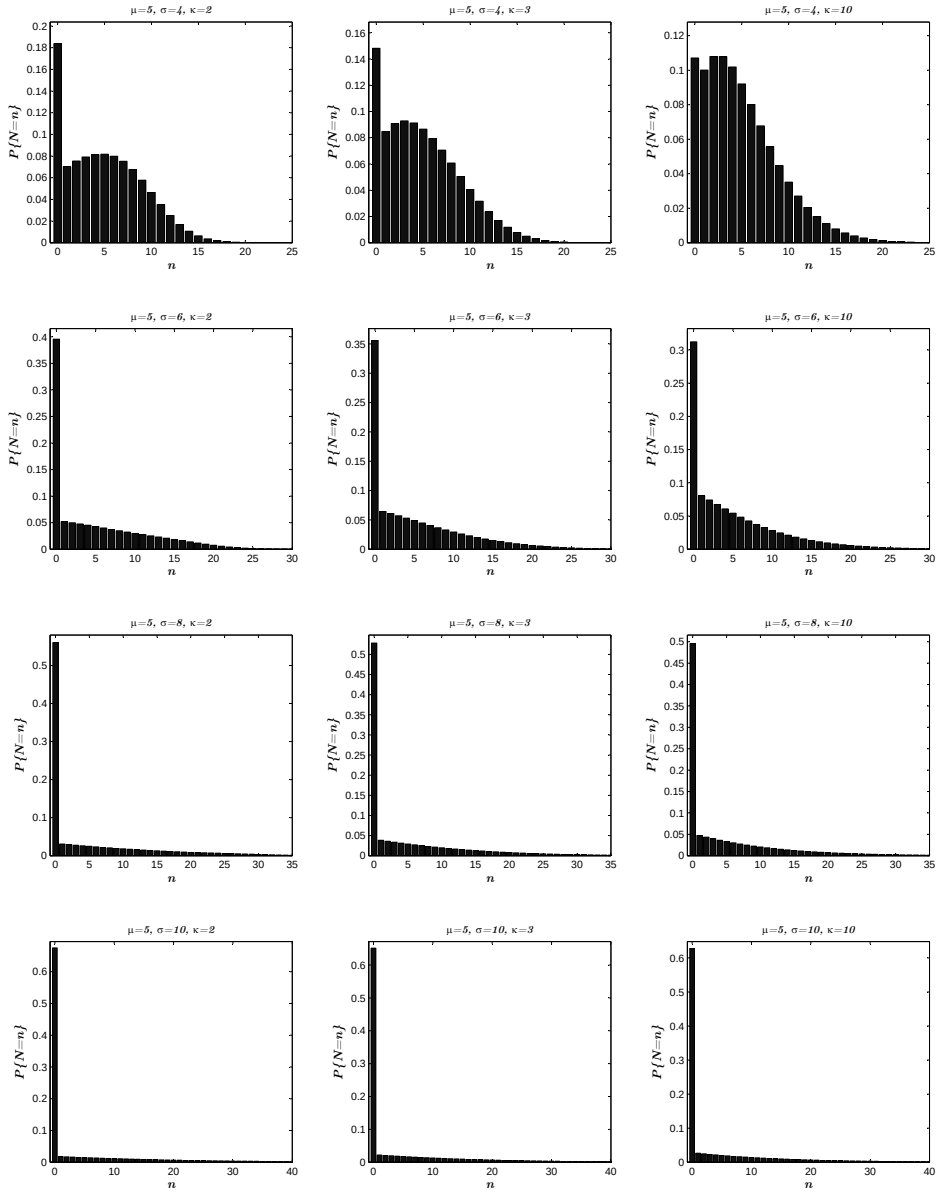

Figure 5.3 Fitted distributions generated by the procedure in Proposition 5.1. Standard deviation and the fitting parameter κ are varied as shown in the plots.

Chapter 6

Base-stock policies for consumables under the use of emergency shipments: State space aggregation and asymptotics

"Slightly wrong equations and identities...

Avogadro's number	:	$69^{\pi^{\sqrt{5}}}$
Gravitational constant G	:	$\frac{1}{e^{(\pi-1)^{(\pi+1)}}}$
R (gas constant)	:	$(e+1)\sqrt{5}$
g	:	$6+\ln(45)$"

xkcd

6.1. Introduction

This chapter studies base-stock policies for consumables that are reviewed periodically. When the stock for consumables is depleted, it is a common procedure to use an emergency supply source to replenish the part almost instantaneously so that maintenance is not halted for lack of a part. All items that are replenished by the

emergency procedure are lost to the normal mode of replenishment. This problem is mathematically equivalent to the classical lost sales inventory problem that has been studied by Karlin and Scarf (1958), Morton (1969), Morton (1971), van Donselaar et al. (1996), Johansen (2001), Janakiraman et al. (2007), Zipkin (2008b), Zipkin (2008a), Levi et al. (2008), Huh et al. (2009b), and Goldberg et al. (2012). This system consists of a periodically reviewed stock point which faces stochastic i.i.d. demand. When demand in a period exceeds the on hand inventory, the excess is lost. Replenishment orders arrive after a lead time τ. At the end of each period, costs for lost sales and holding inventory are charged. For such systems, we are interested in minimizing the long run average cost per period.

The loss of excess demand can have many interpretations. In the context of this thesis, the excess is lost because it is filled by an emergency shipment from another supplier. The traditional interpretation is that a potential sale to a customer has been lost. The terminology of a lost sale has become the standard in the literature so we will also use this standard in the present chapter.

The structure of the optimal policy for lost sales inventory systems with a positive replenishment lead time is still not completely understood, and the computation of optimal policies suffers from the curse of dimensionality as the state space is τ-dimensional. Goldberg et al. (2012) show that the policy to order the same quantity each period is asymptotically optimal as τ approaches infinity. However, for moderate values of τ as encountered in practice, it is difficult to find a good policy. The only policy with a strict performance bound is the dual-balancing policy proposed by Levi et al. (2008). This policy has a cost of no more than twice the optimal costs. In a numerical study, Zipkin (2008a) shows that the dual balancing policy is effective for low per unit lost sales penalty costs, but that base-stock policies perform better in general, especially for high penalty costs. Huh et al. (2009b) show that in fact, base-stock policies are asymptotically optimal as the lost sales penalty costs approach infinity. However, computing the best base-stock policy for a lost-sales inventory problem efficiently remains a challenge. Huh et al. (2009a), p. 398, observe that: "Although base-stock policies have been shown to perform reasonably well in lost sales systems, finding the best base-stock policy, in general, cannot be accomplished analytically and involves simulation optimization techniques". Although the burden of optimization is alleviated by the fact that the average cost under a base-stock policy is convex in the base-stock level (Downs et al., 2001; Janakiraman and Roundy, 2004), evaluating the performance of any given base-stock policy requires either value iteration or simulation.

In this chapter, we provide an efficient method to compute near optimal base-stock levels for lost sales inventory models as well as accurate approximations for the costs of base-stock policies. This method is based on a different view of the dynamics of

a lost sales inventory system, inspired by a relation to the dual sourcing inventory system. This relation has been studied by Sheopuri et al. (2010), and allows us to use ideas similar to those of Arts et al. (2011) for dual-sourcing inventory systems in the context of lost sales inventory systems. Somewhat counter-intuitively, our approach involves moving from a τ-dimensional state space description to a $(\tau+1)$-dimensional state space description, where τ is the order replenishment lead time. This $(\tau+1)$-dimensional state space is the pipeline of all outstanding orders, but not the on-hand inventory. The next key idea to this approach is to aggregate this pipeline of outstanding orders into a single state variable. This is essential to lending tractability as the size of the original state space grows exponentially in both the lead time *and* the base-stock level. By contrast, the aggregated state space grows linearly in the base-stock level only.

From the distribution of this single aggregated state variable, all relevant performance measures can be computed. The distribution of this single state variable can be studied via a Markov chain. For the transition probabilities of this Markov chain, we derive limiting results and show that for the most commonly used demand distributions, the rate of convergence for these limits is at least exponential. We also show that these limiting results satisfy a type of flow conservation property. This flow conservation property relates the average size of an order entering or leaving the pipeline to the total number of items in the pipeline. Based on these results, evaluating a single base-stock policy approximately is as easy as solving $S+1$ linear equations, where S is the base-stock level. Numerical experiments indicate that this approach yields excellent results. Across a test bed that is an extension of the test beds considered by Huh et al. (2009b) and Zipkin (2008a), we find that our approach has cost differences with the best base-stock policy of at most 1.30% and 0.01% on average.

This chapter is organized as follows. The model and notation are described in §6.2. In §6.3-6.5, we analyze the model by aggregating the state space, providing asymptotics for this aggregation and studying the rate of convergence. In §6.6, we define and study flow conservation properties of approximations and verify that our approximation has this property. We consider a few small extensions in §6.7 and give numerical results for our approximation in §6.8. Concluding remarks are provided in §6.9.

6.2. Model

We consider a periodic review single stage inventory system with a replenishment lead time of τ periods ($\tau \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$). Periods are numbered forward in time and demand in period t is denoted D_t and $\{D_t\}_{t=0}^{\infty}$ is a sequence of non-negative i.i.d.

discrete random variables with $0 < \mathbb{E}[D_t] < \infty$. We let D denote the generic single period demand random variable and we let $D^{(k)}$ denote demand over k periods. We denote the order placed in period t by Q_t and note that this order arrives in period $t+\tau$. The pipeline of orders is denoted $\mathbf{Q}_t = (Q_t, Q_{t-1}, \ldots, Q_{t-\tau})$. We let I_t denote the on-hand inventory at the beginning of period t *before* $Q_{t-\tau}$ arrives. The lost sales in period t are denoted by $L_t = (D_t - I_t + Q_{t-\tau})^+$, where $x^+ = \max(0, x)$. In each period, a holding cost of h per unit on-hand inventory *before* the arrival of an order is incurred. Lost sales are penalized with p per lost sale. The system is operated by a base-stock policy with base-stock level $S \in \mathbb{N}_0$. Thus, at the beginning of period t, an order is placed to raise the inventory position Y_t (on-hand inventory plus all outstanding orders) up to the base-stock level S:

$$Q_t = S - Y_t, \tag{6.1}$$

where

$$Y_t = I_t + \sum_{k=t-\tau}^{t-1} Q_k, \quad t \geq 0. \tag{6.2}$$

We assume without loss of generality that $I_0 \leq S$ and $Q_t = 0$ for $t = -\tau, \ldots, -1$, so that $Q_t \geq 0$ for all $t \in \mathbb{N}_0$. The random variable Q_t depends on S; to stress this, we will sometimes use the notation $Q_t(S)$. For each of the variables described, we use the subscript ∞ to denote a random variable in steady state; for instance $\mathbb{P}(I_\infty = x) = \lim_{t\to\infty} \mathbb{P}(I_t = x)$. Some care needs to be taken to ensure steady state variables do exist; Huh et al. (2009a, Theorem 3) prove that a sufficient condition for these steady state random variables to be well defined is $\mathbb{P}(D \leq S/(\tau+1)) > 0$. Most discrete distributions commonly used, such as Poisson, geometric, and (negative) binomial all satisfy this condition. Also any demand distribution with $\mathbb{P}(D = 0) > 0$ verifies this condition. Our objective will be to minimize the long run average cost per period $C(S)$ over the base-stock level S:

$$C(S) = p\mathbb{E}[L_\infty] + h\mathbb{E}[I_\infty]. \tag{6.3}$$

We note that this description of the problem is slightly different from most descriptions in that we account for holding costs at the beginning of a period *before* the order that is due in that period arrives, whereas we account for lost sales at the end of a period. Obviously this convention does not change the long run expected cost per period, but in the analysis, it will make the equations more transparent.

6.3. State space aggregation

The dynamics of I_t, L_t and Q_t are given by

$$I_{t+1} = (I_t + Q_{t-\tau} - D_t)^+, \tag{6.4}$$

$$L_t = (D_t - I_t - Q_{t-\tau})^+, \tag{6.5}$$

$$Q_{t+1} = D_t - L_t. \tag{6.6}$$

Define the pipeline sum, A_t, as the sum of all outstanding orders at time t, including the order that arrives in period t and the order that was placed in period t:

$$A_t = \sum_{k=t-\tau}^{t} Q_k = \mathbf{Q}_t \mathbf{e}^{\mathrm{T}}, \tag{6.7}$$

where $\mathbf{e}$ is the vector of all ones of length $\tau + 1$. For the pipeline sum, we have the following result.

Lemma 6.1 *The following equations hold for all* $t \geq 0$

(a) $A_t + I_t = S$

(b) $A_{t+1} = \min(S, A_t - Q_{t-\tau} + D_t)$

Proof: For (a), we can simply write using (6.1) and (6.2)

$$A_t + I_t = Q_t + \sum_{k=t-\tau}^{t-1} Q_k + I_t = S - Y_t + Y_t = S.$$

For (b), we have

$$\begin{aligned} A_{t+1} &= S - I_{t+1} \\ &= S - (I_t + Q_{t-\tau} - D_t)^+ \\ &= S - (S - A_t + Q_{t-\tau} - D_t)^+ \\ &= \min(S, A_t - Q_{t-\tau} + D_t), \end{aligned}$$

where the first equality follows from part (a), the second by substituting Equation (6.4), the third applying (a) again, and the final equality is easily verified by distinguishing the case $(S - A_t + Q_{t-\tau} - D_t)^+ = 0$ and $(S - A_t + Q_{t-\tau} - D_t)^+ = S - A_t + Q_{t-\tau} - D_t$. □

Finding $\mathbb{E}[A_\infty]$ gives us all the information we need to evaluate $C(S)$ because

$$\mathbb{E}[I_\infty] = S - \mathbb{E}[A_\infty] \tag{6.8}$$

by Lemma 6.1 (a), and

$$\mathbb{E}[L_\infty] = \mathbb{E}[D_\infty] - \mathbb{E}[Q_\infty] = \mathbb{E}[D_\infty] - \mathbb{E}[A_\infty]/(\tau+1) \tag{6.9}$$

by using equations (6.7) and (6.5), and so

$$C(S) = -(h + p/(\tau+1))\mathbb{E}[A_\infty] + hS + p\mathbb{E}[D_\infty]. \tag{6.10}$$

Finally, we note that Lemma 6.1 (b) gives us the basis for a one-dimensional Markov chain for A_t from which we can determine the distribution and mean of A_∞. This Markov chain has transition probabilities $p_{ij} = \mathbb{P}(A_{t+1} = j | A_t = i)$ that can be found by conditioning:

$$p_{ij} = \begin{cases} \lim_{t\to\infty} \sum_{k=0}^{j} \mathbb{P}(Q_{t-\tau} = i + k - j | A_t = i)\mathbb{P}(D_t = k), & \text{if } 0 \le j < S; \\ \lim_{t\to\infty} \sum_{k=0}^{i} \mathbb{P}(Q_{t-\tau} = k | A_t = i)\mathbb{P}(D_t \ge S + k - i), & \text{if } j = S. \end{cases} \tag{6.11}$$

Unfortunately, to evaluate $\lim_{t\to\infty} \mathbb{P}(Q_{t-\tau} = i | A_t = j)$, we need to evaluate the $(\tau+1)$-dimensional Markov chain $\mathbf{Q}_t$. That is,

$$\lim_{t\to\infty} \mathbb{P}\left(Q_{t-\tau} = x | A_t = y\right) = \lim_{t\to\infty} \frac{\sum_{\mathbf{q}|\mathbf{q}_{\tau+1}=x \cap \mathbf{q}\mathbf{e}^{\mathrm{T}}=y} \mathbb{P}(\mathbf{Q}_t = \mathbf{q})}{\sum_{\mathbf{q}|\mathbf{q}\mathbf{e}^{\mathrm{T}}=y} \mathbb{P}(\mathbf{Q}_t = \mathbf{q})}. \tag{6.12}$$

Thus, in this view of the problem, the dimension of the system just increased from τ-dimensional space to $(\tau+1)$-dimensional space and so this task suffers from the curse of dimensionality even more than finding optimal policies does. In fact, it can be shown that the state space of $\mathbf{Q}_t$ grows exponentially in both S and τ as $\binom{S+\tau+1}{S}$. (For a derivation of this result, see §6.A.3.) However, in the limit that $S \to \infty$, we can characterize $\mathbb{P}(Q_{t-\tau} = i | A_t = j)$ using limiting results and we pursue this in the next section.

6.4. Asymptotics

In this section, we show that as S approaches infinity and all other parameters stay constant, that

$$\mathbb{P}(Q_{t-\tau} = i | A_t = j) \to \mathbb{P}\left(D_{t-\tau-1} = i \middle| \sum_{k=t-\tau-1}^{t-1} D_k = j\right). \tag{6.13}$$

Furthermore, for $S = 0, 1$, (6.13) holds with equality in the limit that $t \to \infty$. We use these results to find an asymptotic approximation for $C(S)$. To state our results, we need some additional notation. We let $\xrightarrow{\mathcal{P}}$ denote convergence in probability.[1]

[1] A sequence of random variables X_n is said to converge in probability to X (notation $X_n \xrightarrow{\mathcal{P}} X$) if $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \ge \varepsilon) = 0$ for all $\varepsilon > 0$.

Theorem 6.1 *The following holds for all $t \geq \tau$ when everything is held constant except S:*

(a) As $S \to \infty$, $Q_{t+1} \xrightarrow{\mathcal{P}} D_t$

(b) As $S \to \infty$, $\mathbb{P}(Q_{t+1} = i) \to \mathbb{P}(D_t = i)$.

(c) As $S \to \infty$, $\mathbb{P}(Q_{t-\tau} = i | A_t = j) \to \mathbb{P}\left(D_{t-\tau-1} = i \middle| \sum_{k=t-\tau-1}^{t-1} D_k = j\right)$.

(d) For $S = 0$ and $S = 1$ and $i \leq j \leq S$,

$$\lim_{t\to\infty} \mathbb{P}(Q_{t-\tau} = i | A_t = j) = \mathbb{P}\left(D_{t-\tau-1} = i \middle| \sum_{k=t-\tau-1}^{t-1} D_k = j\right).$$

PROOF: First note that, by Equation (6.6), $Q_{t+1} \leq D_t$ with probability 1 for all $t \geq 0$. This implies in particular that $Q_{t+1} \leq_{\text{st}} D_t$, i.e., $\mathbb{P}(Q_{t+1} \leq x) \geq \mathbb{P}(D_t \leq x)$ and so also

$$\mathbb{P}(A_t \leq x) \geq \mathbb{P}\left(\sum_{k=t-\tau-1}^{t-1} D_k \leq x\right). \tag{6.14}$$

Second, we observe that $Q_{t+1} = D_t$ if and only if $L_t = 0$ which, by Equations (6.5) and Lemma 6.1 (a), is equivalent to the inequality

$$D_t \leq S - A_t + Q_{t-\tau}. \tag{6.15}$$

With this set up, we will now show that as $S \to \infty$, $Q_{t+1} \xrightarrow{\mathcal{P}} D_t$. Let $\delta \in (0, 1)$ and let S_δ satisfy $\mathbb{P}(D^{(\tau+2)} \leq S_\delta) > 1 - \delta$. (Such an $S_\delta < \infty$ exists because $\mathbb{E}[D] < \infty$ and so $\lim_{x\to\infty} \mathbb{P}\left(D^{(\tau+2)} \leq x\right) = 1$.) Now for $S \geq S_\delta$, we have

$$\begin{aligned}
\mathbb{P}(|D_t - Q_{t+1}| > 0) &= \mathbb{P}(D_t - Q_{t+1} > 0) \\
&= 1 - \mathbb{P}(D_t = Q_{t+1}) \\
&= 1 - \mathbb{P}(D_t \leq S - A_t + Q_{t-\tau}) \\
&\leq 1 - \mathbb{P}(D_t + A_t \leq S) \\
&\leq 1 - \mathbb{P}\left(D^{(\tau+2)} \leq S\right) \\
&< 1 - (1 - \delta) = \delta.
\end{aligned} \tag{6.16}$$

The first equality holds because $D_t \geq Q_{t+1}$ with probability one. The second equality holds because D_t and Q_{t+1} are discrete random variables. The third equality holds because, as observed above, $Q_{t+1} = D_t$ if and only if (6.15) holds. The second inequality follows by substituting (6.14), and the final inequality follows from the fact that $S > S_\delta$. This convergence in probability implies also the convergence in distribution asserted in part(b): In the limit that S approaches infinity, $Q_{t+1} \stackrel{d}{=} D_t$ for all $t > \tau$ where $\stackrel{d}{=}$ denotes equality in distribution.

Part (c) now follows from part (b).

For part (d), the case $S = 0$ is trivial. Consider the case $S = 1$. For the condition $A_t = 0$, the result is again trivial. For the condition $S = 1$, we know that at time t, $Q_k = 1$ for exactly one $k \in \{t - \tau, ..., t\}$ and 0 otherwise, because $A_t \leq S$. Thus, the state space of the pipeline $\mathbf{Q}_t$, consists of the zero vector $\mathbf{0}$ and the unit vectors $\mathbf{e}_i$, for $i = 1, \ldots, m$, where $\mathbf{e}_i$ corresponds to the state that $Q_{t+1-i} = 1$ and $Q_k = 0$ if $k \neq t + 1 - i$ and $\mathbf{0}$ corresponds to an empty pipeline. The transition probabilities of $\mathbf{Q}_t$ are given by:

$$\mathbb{P}(\mathbf{Q}_{t+1} = \mathbf{x}|\mathbf{Q} = \mathbf{y}) = \begin{cases} \mathbb{P}(D = 0), & \text{if } \mathbf{x} = \mathbf{0} \text{ and } \mathbf{y} \in \{\mathbf{0}, \mathbf{e}_{\tau+1}\}; \\ \mathbb{P}(D > 0), & \text{if } \mathbf{x} = \mathbf{e}_1 \text{ and } \mathbf{y} \in \{\mathbf{0}, \mathbf{e}_{\tau+1}\}; \\ 1, & \text{if } \mathbf{x} = \mathbf{e}_{i+1} \text{ and } \mathbf{y} = \mathbf{e}_i \text{ for } i \in \{1, \ldots, \tau\}; \\ 0, & \text{otherwise.} \end{cases} \tag{6.17}$$

It is easily verified that the stationary distribution of $\mathbf{Q}_t$ exists and satisfies $\mathbb{P}(\mathbf{Q}_\infty = \mathbf{e}_i) = \mathbb{P}(\mathbf{Q}_\infty = \mathbf{e}_{i+1})$ for $i = 1, \ldots, \tau$. From this, it follows using (6.12) that $\lim_{t\to\infty} \mathbb{P}(Q_{t-\tau} = 1|A_t = j) = \frac{1}{\tau+1}$, and $\mathbb{P}(Q_{t-\tau} = 0|A_t = j) = \frac{\tau}{\tau+1}$. Now, we find

$$\begin{aligned} \mathbb{P}\left(D_{t-\tau-1} = 1 \middle| \textstyle\sum_{k=t-\tau-1}^{t-1} D_k = 1\right) &= \frac{\mathbb{P}\left(D = 1\right)\mathbb{P}\left(D^{(\tau)} = 0\right)}{\mathbb{P}\left(D^{(\tau+1)} = 1\right)} \\ &= \frac{\mathbb{P}\left(D = 1\right)\mathbb{P}\left(D = 0\right)^\tau}{(\tau+1)\mathbb{P}(D = 1)\mathbb{P}\left(D = 0\right)^\tau} \\ &= 1/(\tau + 1). \end{aligned} \tag{6.18}$$

The complement then equals $\tau/(\tau + 1)$. □

To state our next result, we let $\tilde{A}_\infty$ denote the random variable that results from approximating $\mathbb{P}(A_{t+1} = j|A_t = i)$ with limiting results in Theorem 6.1, i.e., $\mathbb{P}\left(\tilde{A}_\infty = x\right) = \tilde{\pi}(x)$ where $\tilde{\pi}(x)$ solves the set of linear equations

$$\tilde{\pi}(j) = \sum_{i=0}^{S} \tilde{\pi}(i)\tilde{p}_{ij}, \quad j = 0, \ldots, S - 1, \qquad \sum_{i=0}^{S} \tilde{\pi}(i) = 1, \tag{6.19}$$

with

$$\tilde{p}_{ij} = \begin{cases} \sum_{k=0}^{j} \mathbb{P}\left(D_{t-\tau-1} = i + k - j \middle| \sum_{k=t-\tau-1}^{t-1} D_k = i\right) \mathbb{P}(D_t = k), & \text{if } j < S; \\ \sum_{k=0}^{i} \mathbb{P}\left(D_{t-\tau-1} = k \middle| \sum_{k=t-\tau-1}^{t-1} D_k = i\right) \mathbb{P}(D_t \geq S + k - i), & \text{if } j = S. \end{cases} \tag{6.20}$$

Furthermore, we let

$$\tilde{C}(S) = -(h + p/(\tau + 1))\mathbb{E}[\tilde{A}_\infty] + hS + p\mathbb{E}\left[D_\infty\right],$$

and $\tilde{I}_\infty = S - \tilde{A}_\infty$ so that $\mathbb{P}(\tilde{I}_\infty = x) = \tilde{\pi}(S - x)$ (by Lemma 6.1 (a)).

Theorem 6.2 *If $\mathbb{P}(D \leq S/(\tau+1)) > 0$, then as $S \to \infty$,*

(a) $\tilde{p}_{ij} \longrightarrow p_{ij}$,

(b) $\tilde{\pi}(x) \longrightarrow \mathbb{P}(A_\infty = x)$,

(c) $\mathbb{E}\left[\tilde{A}_\infty\right] \longrightarrow \mathbb{E}\left[A_\infty\right]$,

(d) $\tilde{C}(S) \longrightarrow C(S)$.

Furthermore we have that $\tilde{C}(1) = C(1)$ and if $\tau = 0$, then $\tilde{C}(S) = C(S)$.

Proof: Part (a) follows directly from Theorem 6.1 (c). From Huh et al. (2009a) Theorem 3, we know that under the condition $\mathbb{P}(D \leq S/(\tau+1)) > 0$, A_∞ is well defined. Consequently, $\mathbb{P}(A_\infty = x)$, $\mathbb{E}[A_\infty]$ and $C(S)$ can all be computed using only algebraic manipulations on $\lim_{t\to\infty} \mathbb{P}(Q_{t-\tau} = i | A_t = j)$. Since limits are preserved under such manipulation, we obtain (b)-(d). That $\tilde{C}(1) = C(1)$ follows from Theorem 6.1 (d), and $\tilde{C}(S) = C(S)$ if $\tau = 0$ follows from observing that A_t is one-dimensional in this case and so $\tilde{A}_t = A_t$ with probability one. □

Even for rather small S, the distributions of I_∞ and A_∞ are very well approximated by the distributions of $\tilde{I}_\infty$ and $\tilde{A}_\infty$. Figure 6.1 illustrates this for I_∞ by showing the distribution of I_∞ as determined by simulation in conjunction with the distribution of $\tilde{I}_\infty$. The same also holds for $\tilde{C}(S)$ compared with $C(S)$ as shown in Figure 6.2.

In §6.8, we report a more elaborate numerical study that shows that the approximations obtained are indeed very good across a much wider range of instances.

We conclude this section by remarking that the results above can be used to efficiently find good base-stock levels for lost sales systems. From Downs et al. (2001), we know that $C(S)$ is convex in S, so a simple heuristic to find a good base-stock level is simply to perform a golden section search (or any other algorithm of choice) on $\tilde{C}(S)$ with the upper bound S_{UB} and lower bound S_{LB} on S given by the result of Huh et al. (2009b):

$$S_{UB} = \inf\left\{y : \mathbb{P}\left(D^{(\tau+1)} \leq y\right) \geq \frac{p/(\tau+1)}{p/(\tau+1)+h}\right\},$$
$$S_{LB} = \inf\left\{y : \mathbb{P}\left(D^{(\tau+1)} \leq y\right) \geq \frac{p+\tau h}{p+(\tau+1)h}\right\}.$$

We call this heuristic the $\mathcal{L}$-heuristic because it is based on limiting results. In the numerical section, we explore this and find that this heuristic is both extremely effective and efficient.

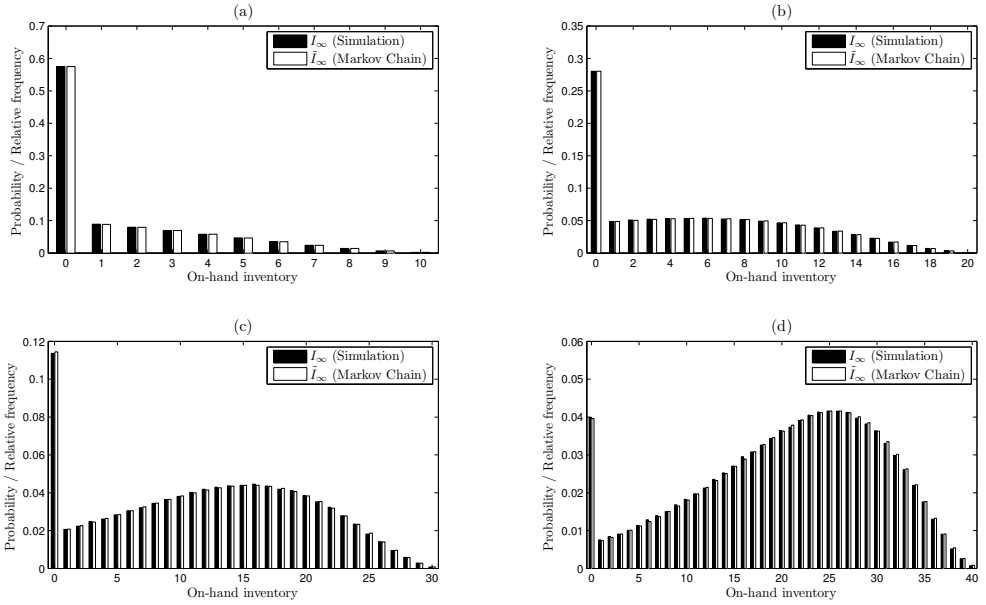

Figure 6.1 The distributions I_∞ as determined by simulation and of $\tilde{I}_\infty$ as determined by solving (6.19) for a lost sales system with lead time $\tau = 4$ facing Geometric demand with mean 5 and base-stock levels of 10, 20, 40 and 40 in (a)-(d) respectively.

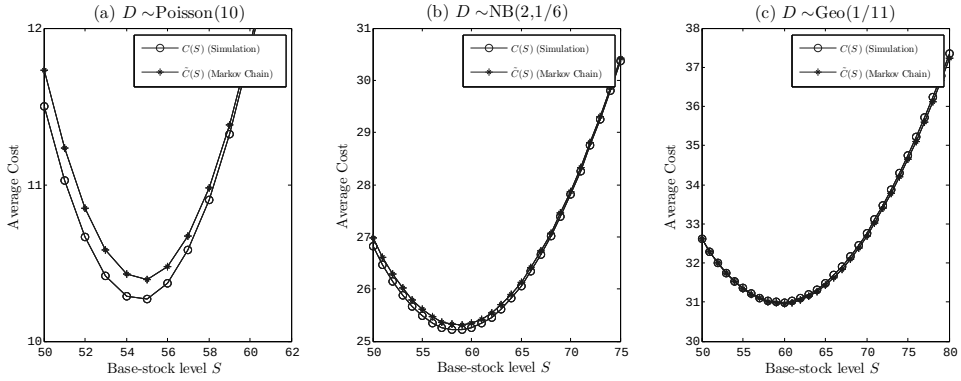

Figure 6.2 The true cost function $C(S)$ and the approximated cost function $\tilde{C}(S)$ for the lost sales system with $\tau = 4$, $h = 1$, $p = 10$ for Poisson, negative binomial and geometric demand in (a)-(c) respectively. The mean demand for all these distributions is 10.

6.5. Rates of convergence

In this section, we show that the asymptotics of the previous section have very good convergence properties under mild conditions on the demand distribution. To state our results we introduce the hazard of demand over k periods as

$$H^{(k)}(x) = \mathbb{P}\left(D^{(k)} = x \middle| D^{(k)} \geq x\right) = \frac{\mathbb{P}\left(D^{(k)} = x\right)}{\mathbb{P}\left(D^{(k)} D_t \geq x\right)}.$$

Oddly, the hazard rate properties of common discrete random variables are not found in standard literature. For the most commonly used demand models, namely Poisson, geometric, and negative binomial, we summarize results in Proposition 6.1.

Proposition 6.1 *If D is a Poisson distributed random variable, then for any $k \in \mathbb{N}$*

$$\liminf_{x\to\infty} H^{(1)}(x) = \lim_{x\to\infty} H^{(1)}(x) = \lim_{x\to\infty} H^{(k)}(x) = \liminf_{x\to\infty} H^{(k)}(x) = 1.$$

Furthermore, if D is a negative binomially (geometrically) distributed random variable with success probability p and r required successes, then for any $k \in \mathbb{N}$

$$\liminf_{x\to\infty} H^{(1)}(x) = \lim_{x\to\infty} H^{(1)}(x) = \lim_{x\to\infty} H^{(k)}(x) = \liminf_{x\to\infty} H^{(k)}(x) = p.$$

The proof of this proposition is in the appendix. With these results, we now turn to the rate of convergence of the limits in §6.4.

Theorem 6.3 *If $\liminf_{x\to\infty} H^{(\tau+2)}(x) = 1 - \theta \in (0,1)$, then Q_{t+1} converges to D_t in probability at least exponentially in S, i.e., for any $\varepsilon \in (0, 1-\theta)$,*

$$\mathbb{P}(D_t - Q_{t+1}(S) > 0) \leq O\left((\theta+\varepsilon)^S\right).$$

Furthermore, if $\liminf_{x\to\infty} H^{(\tau+2)}(x) = 1$, Q_{t+1} converges to D_t in probability super-exponentially in S, i.e., for any $\varepsilon \in (0,1)$,

$$\mathbb{P}(D_t - Q_{t+1}(S) > 0) \leq O\left(\varepsilon^S\right).$$

Proof: From (6.16), we know that $\mathbb{P}(D_t - Q_{t+1}(S) > 0) \leq \mathbb{P}(D^{(\tau+2)} > S)$. Let $\liminf_{x\to\infty} H^{(\tau+2)}(x) = 1 - \theta \in (0,1)$. This implies that for any $\varepsilon \in (0,\theta)$, we can choose an $N \in \mathbb{N}$ such that for all $x > N$, $H^{\tau+2}(x) > 1 - \theta - \varepsilon$. Now fix $C > 0$ such that

$$\mathbb{P}(D^{(\tau+2)} > S) \leq C(\theta+\varepsilon)^S \tag{6.21}$$

for all $S \leq N$. Next observe that for $S \geq N$

$$\begin{aligned}\frac{\mathbb{P}\left(D^{(\tau+2)} > S+1\right)}{\mathbb{P}\left(D^{(\tau+2)} > S\right)} &= \frac{\mathbb{P}\left(D^{(\tau+2)} > S\right) - \mathbb{P}\left(D^{(\tau+2)} = S+1\right)}{\mathbb{P}\left(D^{(\tau+2)} > S\right)} \\ &= 1 - H^{(\tau+2)}(S+1) \leq \theta + \varepsilon. \end{aligned} \tag{6.22}$$

Now we proceed by induction to show that $\mathbb{P}(D^{\tau+2} > S) \leq C(\theta+\varepsilon)^S$ for all $S \in \mathbb{N}$. We have already verified the induction hypothesis that $\mathbb{P}(D^{\tau+2} > S) \leq C(\theta+\varepsilon)^S$ for all $S \leq N$. Suppose it holds for some $S \geq N$ and consider $S+1$:

$$\begin{aligned}\mathbb{P}\left(D^{(\tau+2)} > S+1\right) &= \frac{\mathbb{P}\left(D^{(\tau+2)} > S+1\right)}{\mathbb{P}\left(D^{(\tau+2)} > S\right)}\mathbb{P}\left(D^{(\tau+2)} > S\right) \\ &\leq (\theta+\varepsilon)\mathbb{P}\left(D^{(\tau+2)} > S\right) \\ &\leq (\theta+\varepsilon)C(\theta+\varepsilon)^S = C(\theta+\varepsilon)^{S+1}.\end{aligned}$$

The first inequality holds by using (6.22) and the second follows from the induction hypothesis.

The second part of the proof follows an analogous argument where $\theta = 0$, and so we omit it. □

A direct corollary from combining Proposition 6.1 and Theorem 6.3 is that the rates of convergence asserted in Theorem 6.3 are actually independent of the lead time for the Poisson, negative binomial and geometric demand model.

Corollary 6.1 *If D has a Poisson distribution, then $\mathbb{P}(D_t - Q_{t+1}(S) > 0) \leq O(\varepsilon^S)$ for any $\varepsilon > 0$, regardless of the lead time, τ. If D has a geometric or negative binomial distribution with succes probability p, then $\mathbb{P}(D_t - Q_{t+1}(S) > 0) \leq O(p+\varepsilon^S)$ for any $\varepsilon > 0$, regardless of the lead time, τ.*

Theorem 6.4 *If $\liminf_{x\to\infty} H_D^{(\tau+2)}(x) = 1-\theta \in (0,1]$, $\tilde{A}_\infty$ converges in distribution to A_∞ at least exponentially fast in S, i.e. for any $\varepsilon > 0$ the following hold:*

$$\begin{aligned}\mathbb{P}(A_\infty = x) &= \tilde{\pi}(x) + O\left((\theta+\varepsilon)^S\right), \\ \mathbb{E}\left[A_\infty\right] &= \mathbb{E}[\tilde{A}_\infty] + O\left((\theta+\varepsilon)^S\right), \\ C(S) &= \tilde{C}(S) + O\left((\theta+\varepsilon)^S\right).\end{aligned}$$

The proof of Theorem 6.4 is in §6.A.2. Here too, Proposition 6.1 and Theorem 6.4 can be combined to show that the convergence rate is independent of the lead time for Poisson, negative binomial, and geometric demand.

Corollary 6.2 *If D has a Poisson distribution, then for any $\varepsilon > 0$, it holds that $\mathbb{P}(A_\infty = x) = \tilde{\pi}(x) + O\left(\varepsilon^S\right)$, $\mathbb{E}\left[A_\infty\right] = \mathbb{E}[\tilde{A}_\infty] + O\left(\varepsilon^S\right)$, and $C(S) = \tilde{C}(S) + O\left(\varepsilon^S\right)$. If D has a negative binomial distribution with succes probability p, then for any $\varepsilon > 0$, it holds that $\mathbb{P}(A_\infty = x) = \tilde{\pi}(x) + O\left((p+\varepsilon)^S\right)$, $\mathbb{E}\left[A_\infty\right] = \mathbb{E}[\tilde{A}_\infty] + O\left((p+\varepsilon)^S\right)$, and $C(S) = \tilde{C}(S) + O\left((p+\varepsilon)^S\right)$.*

Since the random variable D is heavy-tailed if and only if, $\lim_{x\to\infty} H^{(1)} = 0$ (Foss et al., 2011), we have no results on the rate of convergence for heavy-tailed demand distributions. However, in the numerical sections we also test our approximation for the heavy-tailed generalized Pareto distribution and find that also here the approximation performs very well.

6.6. Internal consistency: flow conservation

Our approximation relies on aggregating a pipeline of orders into a single state variable. Because A_t is originally a pipeline of orders, everything that goes in has to come out. Furthermore, everything that goes in, stays there for $\tau + 1$ periods. Thus by Little's law, we must have that

$$(\tau + 1)\mathbb{E}[Q_\infty] = \mathbb{E}[A_\infty]. \tag{6.23}$$

Alternatively, we might observe that $A_t = \sum_{k=t-\tau}^{t} Q_k$ also directly implies (6.23). In this light, we may think of (6.23) as expressing flow conservation: Since A_t contains $\tau + 1$ order quantities, on average the outgoing order should equal the total number of items in the pipeline divided by the length of the pipeline. Thus, an attractive property of any approximation of A_t is that it also satisfies (6.23) in some way. Let us make this more precise. Via (6.11), an approximation of

$$\lim_{t\to\infty} \mathbb{P}(Q_{t-\tau} = x | A_t = y)$$

induces an approximate Markov chain for A_t. Let us denote the Markov chain induced by such an approximation $\hat{A}_t$, and let us denote the approximation for $\lim_{t\to\infty} \mathbb{P}(Q_{t-\tau} = x|A_t = y)$ by $\mathbb{P}(\hat{Q}_{t-\tau} = x|\hat{A}_t = y)$. Now under this approximation, the outgoing order has long run mean

$$\mathbb{E}\left[\hat{Q}_\infty\right] = \sum_{y=0}^{S} \mathbb{E}\left[\hat{Q}_{t-\tau}\middle|\hat{A}_t = y\right] \mathbb{P}\left(\hat{A}_\infty = y\right).$$

The next Proposition identifies a large class of approximations $\mathbb{P}(\hat{Q}_{t-\tau} = x|\hat{A} = y)$ that leads to an approximate chain $\hat{A}_t$ that satisfies $(\tau + 1)\mathbb{E}\left[\hat{Q}_\infty\right] = \mathbb{E}\left[\hat{A}_\infty\right]$.

Definition 6.1 A Markov chain $\hat{A}_t$ induced by replacing $\lim_{t\to\infty} \mathbb{P}(Q_{t-\tau} = x|A_t = y)$ with some approximation $\mathbb{P}(\hat{Q}_{t-\tau} = x|\hat{A}_t = y)$ in the transition probabilities (6.11) is called *internally consistent* if it satisfies $(\tau + 1)\mathbb{E}\left[\hat{Q}_\infty\right] = \mathbb{E}\left[\hat{A}_\infty\right]$.

With Definition 6.1 in place, we can state the main result of this section.

Proposition 6.2 *Any Markov chain $\hat{A}_t$ on $0, \ldots, S$ with transition probabilities $\hat{p}_{ij} = \mathbb{P}(\hat{A}_{t+1} = j | \hat{A}_t = i)$ such that*

$$\hat{p}_{ij} = \begin{cases} \sum_{k=0}^{j} \mathbb{P}\left(\hat{Q}_{t-\tau} = i + k - j \middle| \hat{A}_t = i\right) \mathbb{P}(D_t = k), & \text{if } 0 \le j < S; \\ \sum_{k=0}^{i} \mathbb{P}\left(\hat{Q}_{t-\tau} = k \middle| \hat{A}_t = i\right) \mathbb{P}(D_t \ge S + k - i), & \text{if } j = S; \end{cases} \tag{6.24}$$

is internally consistent if

$$\mathbb{P}\left(\hat{Q} = x \middle| \hat{A} = y\right) = \mathbb{P}\left(X_{t-\tau} = x \middle| \sum_{k=t-\tau}^{t} X_k = y\right)$$

for some integer valued non-negative i.i.d. sequence of random variables X_t.

Proof: First observe that $\sum_{x=0}^{y} \mathbb{P}\left(X_{t-\tau} = x \middle| \sum_{k=t-\tau}^{t} X_k = y\right) = 1$ and so $\hat{A}_t$ is a Markov chain indeed.

Now we establish that $(\tau + 1)\mathbb{E}\left[\hat{Q}_\infty\right] = \mathbb{E}\left[\hat{A}_\infty\right]$. Because

$$\mathbb{E}\left[X_n \middle| \sum_{k=t-\tau}^{t} X_k = y\right] = \mathbb{E}\left[X_{n+1} \middle| \sum_{k=t-\tau}^{t} X_k = y\right]$$

for any $n \in \{t - \tau, \ldots, t - 1\}$ and

$$\sum_{n=t-\tau}^{t} \mathbb{E}\left[X_n \middle| \sum_{k=t-\tau}^{t} X_k = y\right] = y,$$

we have that

$$\mathbb{E}\left[X_n \middle| \sum_{k=t-\tau}^{t} X_k = y\right] = y/(\tau + 1). \tag{6.25}$$

Now for $\mathbb{E}\left[\hat{Q}_\infty\right]$ we find

$$\begin{aligned} \mathbb{E}\left[\hat{Q}_\infty\right] &= \sum_{y=0}^{S} \mathbb{E}\left[X_{t-\tau} \middle| \sum_{k=t-\tau}^{t} X_k = y\right] \mathbb{P}\left(\hat{A}_\infty = y\right) \\ &= \sum_{y=0}^{S} y/(\tau + 1) \mathbb{P}\left(\hat{A}_\infty = y\right) \\ &= \mathbb{E}\left[\hat{A}_\infty\right] /(\tau + 1). \end{aligned}$$

The second equality holds by substituting (6.25). □

Of all possible choices for X_t in Proposition 6.2, D_t is of course the most obvious because by Theorem 6.1, $Q_t \xrightarrow{\mathcal{P}} D_{t-1}$.

Corollary 6.3 *$\tilde{A}_t$ is internally consistent.*

PROOF: This follows from Proposition 6.2 and the assumption that D_t is a series of i.i.d. discrete non-negative random variables. □

6.7. Extensions

The results in the previous sections can be used for several variations of lost sales inventory models. Below we discuss several such extensions.

6.7.1 General single period cost functions

Our results give approximations, not only for the moments of I_∞ and L_∞, but also for their entire distribution. Thus, a cost function that is not necessarily linear in I_t and L_t can also be accommodated. To see how the distribution of L_∞ and I_∞ can be approximated by the given results, note that by Lemma 6.1 $\mathbb{P}(I_\infty = x) = \mathbb{P}(A_\infty = S - x)$ and using Theorem 6.2, this can be approximated by $\tilde{\pi}(S - x)$. Furthermore, for the distribution of L_t we have for $x > 0$

$$\begin{aligned}\mathbb{P}(L_t = x) &= \mathbb{P}\left((D_t - I_t - Q_{t-\tau})^+ = x\right) \\ &= \sum_{y=0}^{S} \mathbb{P}\left(D_t = x + y + Q_{t-\tau} | A_t = S - y\right) \mathbb{P}(I_t = y) \\ &= \sum_{y=0}^{S} \sum_{z=0}^{S-y} \mathbb{P}(D_t = x + y + z) \mathbb{P}(I_t = y) \mathbb{P}\left(Q_{t-\tau} = z | A_t = S - y\right). \quad (6.26)\end{aligned}$$

Now letting $t \to \infty$ in (6.26) and using the limit results in Theorems 6.1 and 6.2 to approximate, we find (again for $x > 0$):

$$\begin{aligned}&\mathbb{P}(L_\infty = x) = \\ &\sum_{y=0}^{S} \sum_{z=0}^{S-y} \mathbb{P}(D_t = x + y + z) \tilde{\pi}(S - y) \mathbb{P}\left(D_{t-\tau-1} = z \middle| \textstyle\sum_{k=t-\tau-1}^{t-1} D_k = S - y\right).\end{aligned}$$

6.7.2 Service level constraints

Suppose we are interested in the service level of a lost sales system in that we require that a fraction $\beta \in [0, 1)$ of all demand is filled while minimizing the on-hand inventory.

If we choose to control this system by a base-stock policy, the objective now becomes to minimize S such that

$$\beta \mathbb{E}[D] \leq \mathbb{E}[A_\infty]/(\tau+1). \tag{6.27}$$

An approximate solution to this problem can be found by approximating $\mathbb{E}[A_\infty]$ by $\mathbb{E}\left[\tilde{A}_\infty\right]$.

6.8. Numerical results

We test how good the base-stock policies found by using our limiting results are, compared to the best base-stock policies. We use and extend the test bed of Huh et al. (2009b) which is an extension of the test bed of Zipkin (2008a). (Note that the papers of Zipkin (2008a) and Huh et al. (2009b) also report the performance of the globally optimal replenishment policy.) The first set of instances in this test bed have Poisson or Geometric demand distributions, both with mean 5 and lead times $\tau \in \{1,2,3,4\}$. The holding cost is kept constant at $h=1$ while the penalty costs $p \in \{1,4,9,19,49,99,199\}$. In keeping with how results on the test bed are reported in Zipkin (2008a) and Huh et al. (2009b), the detailed numerical results per instance are reported in Appendix 6.C. In this section, we only present aggregated results about the gap with the best base-stock level and the accuracy of the cost estimates of our approximation (both computed by simulation optimization). We also compare our heuristic against those suggested by Huh et al. (2009b). The first heuristic they suggest is to select the base-stock level that minimizes cost for a backorder system with p as the cost per backorder per period. The resulting base-stock level is denoted $S^{\mathcal{B}}$ and is the solution to a news vendor problem:

$$S^{\mathcal{B}} = \inf\left\{y : \mathbb{P}\left(D^{(\tau+1)} \leq y\right) \geq \frac{p+\tau h}{p+(\tau+1)h}\right\}.$$

We call this heuristic the $\mathcal{B}$-heuristic ($\mathcal{B}$ for backlogging). Generally, it performs quite poorly and so Huh et al. (2009b) also suggest an improved heuristic also based on solving news vendor problems. This improved heuristic has base-stock level $S^{\mathcal{I}}$ that satisfies

$$S^{\mathcal{I}} = \frac{p}{p+h}\inf\left\{y : \mathbb{P}\left(D^{(\tau+1)} \leq y\right) \geq \frac{p}{p+h}\right\} + \frac{h}{p+h}\inf\left\{y : \mathbb{P}\left(D \leq y\right) \geq \frac{p}{p+h}\right\}.$$

We call this heuristic the $\mathcal{I}$-heuristic ($\mathcal{I}$ for improved news vendor heuristic). Our heuristic is to use the base-stock level $S^{\mathcal{L}}$ which is obtained by minimizing $\tilde{C}(S)$ using a golden section search. We call our heuristic the $\mathcal{L}$-heuristic.

Tables 6.1 and 6.2 report the average and (absolute) maximum percentage errors in the performance predicted by the $\mathcal{L}$ heuristic

$$100\% \cdot \left(\tilde{C}(S^{\mathcal{L}}) - C(S^{\mathcal{L}})\right) / C(S^{\mathcal{L}}),$$

the average and maximum percentage cost differences with the best base-stock policy and the hitrate: the percentage of instances in which the $\mathcal{L}$-heuristic finds the best base-stock level. In all cases except one, the $\mathcal{L}$-heuristic finds the optimal base-stock level. In the single case that it does not, the optimality gap is only 1.01%. The estimate of the cost that the $\mathcal{L}$-heuristic provides is extremely accurate for geometric demand and only slightly less so for Poisson demand. The fact that the $\mathcal{L}$-heuristic also provides an accurate approximation of the average cost rate is an asset, because the other heuristics only provide a base-stock level without an (accurate) cost approximation.

Table 6.1 Performance of the $\mathcal{L}$-heuristic for Poisson demand with mean 5

Parameter	Estimation error (%)		Gap with best base-stock (%)		Hitrate (%)
Lead time	AVG	MAX	AVG	MAX	
1	0.99	3.53	0.14	1.01	86
2	1.22	3.69	0.00	0.00	100
3	1.25	3.84	0.00	0.00	100
4	1.18	3.68	0.00	0.00	100
Penalty cost					
1	3.68	3.84	0.00	0.00	100
4	1.78	1.98	0.00	0.00	100
9	0.87	1.28	0.25	1.01	75
19	0.74	0.78	0.00	0.00	100
49	0.45	0.61	0.00	0.00	100
99	0.34	0.42	0.00	0.00	100
199	0.24	0.37	0.00	0.00	100
Total	1.16	3.84	0.04	1.01	96

At an aggregate level we can compare the three heuristics for these instances and this comparison is shown in Table 6.3. For these problem instances, the $\mathcal{L}$ heuristic outperforms the other heuristics with a significant margin.

Next, we look at instances facing Poisson demand and means ranging from 1 to 10. Holding cost h is kept constant again at 1, $\tau \in \{2, 4\}$ and $p \in \{1, 4, 9, 19, 49, 99, 199\}$. Table 6.4 shows results for the $\mathcal{L}$-heuristic, and Table 6.5 compares the three heuristics. The $\mathcal{L}$-heuristic again has favorable performance. The largest percentage difference with the best base-stock policy of 1.30% is found here for the instance with $\tau = 2$, $p = 19$ and a mean demand of 2. The optimal base-stock level is 9, whereas our

Table 6.2 Performance of the $\mathcal{L}$-heuristic for geometric demand with mean 5

Parameter	Estimation error (%)		Gap with best base-stock (%)		Hitrate (%)
Lead time	AVG	MAX	AVG	MAX	
1	0.01	0.09	0.00	0.00	100
2	-0.01	-0.05	0.00	0.00	100
3	0.02	0.11	0.00	0.00	100
4	-0.01	-0.04	0.00	0.00	100
Penalty cost					
1	0.03	0.11	0.00	0.00	100
4	0.00	-0.05	0.00	0.00	100
9	0.00	-0.02	0.00	0.00	100
19	0.00	-0.02	0.00	0.00	100
49	0.01	0.01	0.00	0.00	100
99	0.00	-0.01	0.00	0.00	100
199	0.00	-0.01	0.00	0.00	100
Total	0.00	0.11	0.00	0.00	100

Table 6.3 Comparison of heuristics for Poisson and Geometric demand with mean 5

	Demand			
	Poisson with mean 5		Geometric with mean 5	
	Gap with best base-stock (%)		Gap with best base-stock (%)	
Heuristic	AVG	MAX	AVG	MAX
$\mathcal{L}$	0.04	1.01	0.00	0.00
$\mathcal{I}$	1.14	5.59	4.51	11.22
$\mathcal{B}$	20.16	156.72	30.67	232.25

heuristic finds a base-stock level of 10. The performance estimates of the $\mathcal{L}$-heuristic are not very accurate with errors up to 8.26%.

When we compare the three heuristics for these instances (see Table 6.5), we again observe that the $\mathcal{L}$ heuristic performs better than the other heuristics by a significant margin. The largest gap that we found for the $\mathcal{L}$-heuristic of 1.30% is still small compared to the results of the runner up $\mathcal{I}$-heuristic with average and maximum gaps of 4.51% and 11.22% respectively.

Next we consider negative binomial demand with $r \in \{1, 2\}$ required successes and succes probability $q \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$; see Table 6.6. The performance of our heuristic is very good here, both in terms of difference with the best base-stock policy and as a predictor of the actual costs involved. It is perhaps striking that the performance of the $\mathcal{L}$-heuristic is better for negative binomial demand than it is for Poisson demand, even though the theoretical convergence properties are stronger

Table 6.4 Performance of the $\mathcal{L}$-heuristic for Poisson demand with varying means

Parameter	Estimation error (%)		Gap with best base-stock (%)		Hitrate (%)
Lead time	AVG	MAX	AVG	MAX	
2	1.15	4.44	0.03	1.30	96
4	2.41	8.26	0.01	0.15	93
Penalty cost					
1	3.41	4.58	0.01	0.11	90
4	1.86	2.48	0.01	0.24	95
9	1.30	2.04	0.01	0.15	95
19	0.97	1.97	0.10	1.30	85
49	1.10	2.54	0.00	0.00	100
99	1.47	4.33	0.00	0.06	95
199	2.34	8.26	0.00	0.00	100
Mean demand					
1	1.46	6.62	0.00	0.00	100
2	1.24	3.62	0.10	1.30	86
3	2.18	8.26	0.01	0.11	93
4	1.57	3.65	0.00	0.00	100
5	1.84	4.22	0.00	0.00	100
6	2.00	5.10	0.00	0.00	100
7	2.31	5.97	0.02	0.24	93
8	1.65	4.40	0.00	0.00	100
9	2.01	4.54	0.07	0.74	79
10	1.55	4.52	0.00	0.02	93
Total	1.78	8.26	0.02	1.30	94

Table 6.5 Comparison of different heuristics for Poisson demand with varying mean and lead time $\tau = 2$

	Gap with best base-stock (%)	
Heuristic	AVG	MAX
$\mathcal{L}$	0.02	1.30
$\mathcal{I}$	1.90	19.00
$\mathcal{B}$	15.71	98.65

for Poisson demand; see Theorem 6.3 and Proposition 6.1.

We compare the three heuristics under negative binomial demand only for $\tau = 2$, because this is the only lead time considered in the original test bed of Huh et al. (2009b); see Table 6.7. Here again, the $\mathcal{L}$-heuristic outperforms the other heuristics by a considerable margin.

Finally, we test the $\mathcal{L}$-heuristic for instances where demand follows the heavy-tailed

Table 6.6 Performance of the $\mathcal{L}$-heuristic with negative binomial demand

Parameter	Estimation error (%)		Gap with best base-stock (%)		Hitrate (%)
Lead time	AVG	MAX	AVG	MAX	
2	-0.07	-15.01	0.01	0.17	94
4	0.51	5.05	0.00	0.08	90
Penalty cost					
1	0.33	0.87	0.00	0.06	95
4	0.32	0.72	0.00	0.00	90
9	0.26	0.51	0.01	0.17	90
19	0.27	0.84	0.00	0.02	85
49	0.31	1.85	0.01	0.15	95
99	-0.39	-15.01	0.00	0.08	95
199	0.46	5.05	0.00	0.06	95
Negative Binomial parameters (r, q)					
$(1, 0.1)$	0.75	4.48	0.02	0.17	71
$(1, 0.2)$	-0.29	-2.85	0.00	0.02	93
$(1, 0.3)$	-1.42	-15.01	0.01	0.08	93
$(1, 0.4)$	0.78	4.79	0.00	0.00	100
$(1, 0.5)$	0.83	5.05	0.00	0.00	100
$(2, 0.1)$	0.13	-1.59	0.00	0.01	86
$(2, 0.2)$	0.28	0.87	0.00	0.00	100
$(2, 0.3)$	0.27	0.66	0.01	0.15	79
$(2, 0.4)$	0.71	2.95	0.00	0.00	100
$(2, 0.5)$	0.18	0.44	0.00	0.00	100
Total	0.22	5.05	0.00	0.17	92

Table 6.7 Comparison of different heuristics for negative binomial demand and lead time $\tau = 2$

	Gap with best base stock (%)	
Heuristic	AVG	MAX
$\mathcal{L}$	0.01	0.17
$\mathcal{I}$	3.25	15.49
$\mathcal{B}$	22.45	125.39

discretized generalized Pareto distribution. A brief description of this distribution is given in Appendix 6.B for those not familiar with it. We include tests with this distribution because it is heavy-tailed and so none of the convergence rate results in §6.5 apply. We consider the discretized generalized Pareto distribution with shape parameter $k = 0.1$ and scale parameter $\sigma = 5$ (see Table 6.8) and with shape parameter $k = 0.4$ and scale parameter $\sigma = 10$ (see Table 6.9). Here too, our heuristic performs very good. Perhaps surprisingly, our heuristic identifies better base-stock

Table 6.8 Performance of the $\mathcal{L}$-heuristic with discretized generalized Pareto demand with $k = 0.1$ and $\sigma = 5$

Parameter	Estimation error (%)		Gap with best base-stock (%)		Hitrate
Lead time	AVG	MAX	AVG	MAX	
1	-0.26	-1.95	0.00	0.00	100
2	-0.33	-2.32	0.01	0.06	71
3	-0.41	-3.25	0.00	0.00	100
4	-0.50	-1.02	0.00	0.00	100
Penalty cost					
1	-0.13	-0.21	0.00	0.00	100
4	-0.23	-0.31	0.00	0.00	100
9	-0.31	-0.49	0.00	0.00	100
19	-0.41	-0.66	0.00	0.00	100
49	-0.75	-2.32	0.00	0.01	75
99	-0.49	-1.95	0.00	0.00	100
199	-0.32	-3.25	0.01	0.06	75
Total	-0.37	-3.25	0.00	0.06	93

levels here, than it does for Poisson demand, even though there are theoretically excellent convergence results for Poisson demand. A plausible explanation for this is that for finite S, internal consistency as outlined in §6.6 is more instrumental in the quality of our approximation than the asymptotic results in §6.4. We do see that the hitrate deteriorates significantly as p increases in Table 6.9. This is because in these cases, optimal base-stock levels are high and so the exact optimum is easier to miss.

In closing, we comment on computation times. Evaluating the best base-stock policy using value iteration is almost as difficult as determining the optimal policy. Bijvank and Johansen (2012) use a value iteration algorithm in a very similar setting and report computation times of several minutes op to several hours. We already observed that the state space required to evaluate the performance of a base-stock policy grows exponentially in both S and τ. By contrast, the state space of our approximation grows linearly in S only.

We determined the optimal base-stock levels with simulation and found computation times of several minutes to be the norm on a machine with 2.4 GHz Intel processor and 4GB of RAM. By contrast, the $\mathcal{L}$-heuristic finds near optimal base-stock levels within less then 0.1 seconds on the same machine.

Table 6.9 Performance of the $\mathcal{L}$-heuristic with discretized generalized Pareto demand with $k = 0.4$ and $\sigma = 10$

Parameter	Estimation error (%)		Gap with best base-stock (%)		Hitrate
Lead time	AVG	MAX	AVG	MAX	
1	-0.60	-1.97	0.00	0.00	86
2	-0.29	-0.68	0.00	0.01	43
3	-0.84	-1.75	0.00	0.01	57
4	-0.54	-1.23	0.01	0.02	57
Penalty cost					
1	-0.23	-0.28	0.00	0.00	100
4	-0.66	-0.76	0.00	0.01	50
9	-0.66	-0.72	0.00	0.00	100
19	-0.68	-1.23	0.00	0.01	75
49	-0.92	-1.75	0.00	0.01	50
99	-0.80	-1.97	0.01	0.01	25
199	-0.01	-1.49	0.01	0.02	0
Total	-0.57	-1.97	0.00	0.02	54

6.9. Conclusion

We have found an efficient heuristic to find good base-stock levels for lost sales inventory systems. This heuristic outperforms existing heuristics by a considerable margin and also provides accurate cost estimates. This method is based on state space aggregation and limiting results for transition probabilities within this aggregated state space. Numerical experiments indicate that this method has superior performance in a wide diversity of instances with cost differences with the best base-stock policy of at most 1.30% and 0.01% on average. Furthermore, our heuristic is computationally very efficient, the most demanding algorithmic requirement being the solution of linear equations.

6.A. Proofs

6.A.1 Proof of Proposition 6.1

PROOF: Let μ denote the mean of the Poisson distributed random variable D. Consider $H^{(1)}(x)$:

$$\begin{aligned} H^{(1)}(x) &= \frac{e^{-\mu}\frac{\mu^x}{x!}}{\sum_{k=x}^{\infty} e^{-\mu}\frac{\mu^k}{k!}} = \frac{\frac{\mu^x}{x!}}{\frac{\mu^x}{x!} + \sum_{k=x+1}^{\infty}\frac{\mu^k}{k!}} \\ &= \frac{1}{1 + \frac{x!}{\mu^x}\sum_{k=x+1}^{\infty}\frac{\mu^k}{k!}} = \frac{1}{1 + \sum_{k=1}^{\infty}\frac{\mu^k}{\prod_{j=1}^{k}(x+j)}} \\ &\geq \frac{1}{1 + \sum_{k=1}^{\infty}(\mu/x)^k}. \end{aligned} \tag{6.28}$$

Now using that $\lim_{a\to 0}\sum_{k=1}^{\infty} a^k = \lim_{a\to 0} a/(1-a) = 0$, we observe that (6.28) implies that $\lim_{x\to\infty} H^{(1)}(x) \geq \lim_{x\to\infty}\frac{1}{1+\sum_{k=1}^{\infty}(\mu/x)^k} = 1$. Noting that $H^{(1)}(x) < 1$ for all $x \in \mathbb{N}_0$, we have by the squeeze theorem that $\liminf_{x\to\infty} H^{(1)}(x) = \lim_{x\to\infty} H^{(1)}(x) = 1$. Since the Poisson distribution is closed under convolutions, we also have $\liminf_{x\to\infty} H^{(k)}(x) = \lim_{x\to\infty} H^{(k)}(x) = 1$, for any $k \in \mathbb{N}$.

In case $r = 1$, the second result is trivial because then D is a geometric random variable and $H^{(1)}(x) = p$ for all $x \in \mathbb{N}_0$. Consider now the case $r > 1$.

$$\begin{aligned} H^{(1)}(x) &= \frac{\binom{x+r-1}{x}p^r(1-p)x}{\sum_{k=0}^{\infty}\binom{x+r-1+k}{x+k}p^r(1-p)^{x+k}} \\ &= \frac{\binom{x+r-1}{x}p^r(1-p)x}{\binom{x+r-1}{x}p^r(1-p)x + \sum_{k=1}^{\infty}\binom{x+r-1+k}{x+k}p^r(1-p)^{x+k}} \\ &= \frac{1}{1 + \frac{x!(r-1)!}{(x+r-1)!}\frac{1}{p^r(1-p)^x}\sum_{k=1}^{\infty}\binom{x+r-1+k}{x+k}p^r(1-p)^{x+k}} \end{aligned} \tag{6.29}$$

To take the limit as $x \to \infty$, we will now further simplify the second term in the

denominator:

$$\begin{aligned}
&\frac{x!(r-1)!}{(x+r-1)!}\frac{1}{p^r(1-p)^x}\sum_{k=1}^{\infty}\binom{x+r-1+k}{x+k}p^r(1-p)^{x+k}\\
&\quad=\frac{x!(r-1)!}{(x+r-1)!(1-p)^x}\sum_{k=1}^{\infty}\frac{(x+r-1+k)!}{(x+k)!(r-1)!}(1-p)^{x+k}\\
&\quad=\frac{x!}{(x+r-1)!}\sum_{k=1}^{\infty}\frac{(x+r-1+k)!}{(x+k)!}(1-p)^k\\
&\quad=\sum_{k=1}^{\infty}(1-p)^k\prod_{i=1}^{k}\frac{x+r-1+i}{x+i}
\end{aligned}$$

Now since $\lim_{x\to\infty}\prod_{i=1}^{k}\frac{x+r-1+i}{x+i}=1$ for all $k\in\mathbb{N}$, we have that

$$\lim_{x\to\infty}\sum_{k=1}^{\infty}(1-p)^k\prod_{i=1}^{k}\frac{x+r-1+i}{x+i}=\sum_{k=1}^{\infty}(1-p)^k=\frac{1-p}{p} \tag{6.30}$$

Taking the limit as $x\to\infty$ of (6.29) using (6.30) we find

$$\liminf_{x\to\infty}H^{(1)}(x)=\lim_{x\to\infty}H^{(1)}(x)=\frac{1}{1+\frac{1-p}{p}}=p$$

Next, by observing that the sum of negative binomial (geometric) random variables with the same succes probability p also has a negative binomial distribution with succes probability p, we conclude that $\liminf_{x\to\infty}H^{(k)}(x)=\lim_{x\to\infty}H^{(k)}(x)=p$, for any $k\in\mathbb{N}$. □

6.A.2 Proof of Theorem 6.4

Proof: We prove that the exponential convergence in probability of Q_{t+1} to D_t implies exponential convergence in distribution. The entire theorem then follows, as from then on, only algebraic manipulations are involved. Recall that $Q_{t+1}\leq D_t$ with probability one and so for any $a\in\mathbb{N}_0$

$$\mathbb{P}(D_t\leq a)\leq\mathbb{P}(Q_{t+1}\leq a). \tag{6.31}$$

Now for this same a, we have:

$$\begin{aligned}
\mathbb{P}(Q_{t+1}\leq a)&=\mathbb{P}(Q_{t+1}\leq a\cap D_t\leq a)+\mathbb{P}(Q_{t+1}\leq a\cap D_t>a)\\
&\leq\mathbb{P}(D_t\leq a)+\mathbb{P}(Q_{t+1}-D_t\leq a-D_t\cap a-D_t<0)\\
&\leq\mathbb{P}(D_t\leq a)+\mathbb{P}(D_t-Q_{t+1}>0)\\
&=\mathbb{P}(D_t\leq a)+O\left((\theta+\varepsilon)^S\right),
\end{aligned} \tag{6.32}$$

where (6.32) follows from applying Theorem 6.3. Combining (6.31) and (6.32) yields the desired result. □

6.A.3 Derivation of the state space size of $\mathbf{Q}_t$

The size of the state space of the vector Markov chain $\mathbf{Q}_t$ is

$$\mathfrak{S}(S,\tau) = \left|\left\{\mathbf{x} \in \mathbb{N}_0^{\tau+1} | \mathbf{x}\mathbf{e}^{\mathrm{T}} \leq S\right\}\right|.$$

Now observe that $\mathfrak{S}(S,\tau)$ can be expressed recursively in τ. We have for $\tau = 0$

$$\mathfrak{S}(S,0) = \sum_{k=0}^{S} 1 = S+1. \tag{6.33}$$

For $\tau = 1$ we have similarly

$$\mathfrak{S}(S,1) = \sum_{k_1=0}^{S} \sum_{k_2=0}^{S-k_1} 1 = \frac{1}{2}(S+1)(S+2), \tag{6.34}$$

where the second equality follows from substituting (6.33). We can continue such back substitution to obtain

$$\mathfrak{S}(S,2) = \sum_{k_1=0}^{S} \sum_{k_2=0}^{S-k_1} \sum_{k_3=0}^{S-k_1-k_2} 1 = \frac{1}{6}(S+1)(S+2)(S+3) \tag{6.35}$$

$$\mathfrak{S}(S,3) = \sum_{k_1=0}^{S} \sum_{k_2=0}^{S-k_1} \sum_{k_3=0}^{S-k_1-k_2} \sum_{k_4=0}^{S-k_1-k_2-k_3} 1 = \frac{1}{24}(S+1)(S+2)(S+3)(S+4). \tag{6.36}$$

It is now easy to see that

$$\mathfrak{S}(S,\tau) = \frac{1}{(\tau+1)!}\prod_{k=1}^{\tau+1}(S+k) = \frac{(S+\tau+1)!}{S!(\tau+1)!} = \binom{S+\tau+1}{S}. \tag{6.37}$$

Thus, $\mathfrak{S}$ grows exponentially in both τ and S.

6.B. The generalized Pareto distribution

A non-negative continuous random variable X is said to have a generalized Pareto distribution if

$$\mathbb{P}(X < x) = F(x) = 1 - (1 + kx/\sigma)^{-1/k}$$

for some $k > 0$ (shape parameter), $\sigma > 0$ (scale parameter) and all $x > 0$[2]. If $k < 1$, X has finite mean

$$\mathbb{E}[X] = \sigma/(1-k),$$

and if $k < 1/2$, it also has finite variance

$$\mathbf{Var}[X] = \frac{\sigma^2}{(1-k)^2(1-2k)}.$$

It is easily verified that X has a heavy-tail. If $Y = \lfloor X + 1/2 \rfloor$, then Y is said to have a discretized generalized Pareto distribution and

$$\mathbb{P}(Y = y) = F(y + 1/2) - F(y - 1/2)$$

for $y \in \mathbb{N}$.

6.C. Tables with details per instance

Table 6.10 Performance of Limiting base-stock policy for Poisson demand distribution with mean 5

Lead time	Lost sales penalty	Best base-stock		Limiting base-stock			Estimation Error (%)	Diff. from best base-stock (%)
		Level	Cost	Level	Estimated Cost	Real Cost		
1	1	8	2.08	8	2.15	2.08	3.53	0.00
1	4	12	4.16	12	4.23	4.16	1.74	0.00
1	9	13	5.55	14	5.61	5.61	0.04	1.01
1	19	15	6.73	15	6.78	6.73	0.76	0.00
1	49	17	8.22	17	8.25	8.22	0.32	0.00
1	99	18	9.20	18	9.23	9.20	0.32	0.00
1	199	19	10.14	19	10.16	10.14	0.21	0.00
2	1	12	2.23	12	2.31	2.23	3.69	0.00
2	4	16	4.64	16	4.73	4.64	1.98	0.00
2	9	19	6.32	19	6.38	6.32	1.00	0.00
2	19	21	7.84	21	7.89	7.84	0.68	0.00
2	49	23	9.63	23	9.67	9.63	0.38	0.00
2	99	24	10.84	24	10.89	10.84	0.42	0.00
2	199	25	12.03	25	12.07	12.03	0.37	0.00

Continued on next page

[2]The generalized Pareto distribution can, and sometimes is, generalized further by introducing a location parameter and also allowing $k \leq 0$.

Lead time	Lost sales penalty	Best base-stock		Limiting base-stock			Estimation Error (%)	Diff. from best base-stock (%)
		Level	Cost	Level	Estimated Cost	Real Cost		
3	1	15	2.31	15	2.40	2.31	3.84	0.00
3	4	20	4.98	21	5.06	4.98	1.68	0.00
3	9	23	6.86	24	6.95	6.86	1.28	0.00
3	19	26	8.60	26	8.67	8.60	0.78	0.00
3	49	28	10.73	28	10.80	10.73	0.61	0.00
3	99	30	12.15	30	12.19	12.15	0.34	0.00
3	199	32	13.52	32	13.55	13.52	0.18	0.00
4	1	18	2.37	18	2.46	2.37	3.68	0.00
4	4	25	5.20	25	5.29	5.20	1.71	0.00
4	9	28	7.27	28	7.36	7.27	1.19	0.00
4	19	31	9.23	31	9.30	9.23	0.75	0.00
4	49	34	11.60	34	11.66	11.60	0.48	0.00
4	99	36	13.24	36	13.28	13.24	0.27	0.00
4	199	38	14.77	38	14.80	14.77	0.18	0.00

End of Table

Table 6.11 Performance of Limiting base-stock policy for Geometric demand distribution with mean 5

Lead time	Lost sales penalty	Best base-stock		Limiting base-stock			Estimation Error (%)	Diff. from best base-stock (%)
		Level	Cost	Level	Estimated Cost	Real Cost		
1	1	5	4.06	5	4.06	4.06	0.09	0.00
1	4	12	10.04	12	10.04	10.04	0.02	0.00
1	9	17	14.73	17	14.73	14.73	-0.02	0.00
1	19	22	19.40	22	19.40	19.40	0.01	0.00
1	49	29	25.47	29	25.47	25.47	0.01	0.00
1	99	33	29.99	33	29.99	29.99	0.00	0.00
1	199	38	34.41	38	34.41	34.41	-0.01	0.00
2	1	6	4.18	6	4.18	4.18	-0.05	0.00
2	4	15	10.71	15	10.70	10.71	-0.05	0.00
2	9	22	15.99	22	15.99	15.99	-0.01	0.00
2	19	28	21.31	28	21.31	21.31	-0.01	0.00
2	49	36	28.22	36	28.22	28.22	0.01	0.00
2	99	41	33.28	41	33.28	33.28	0.00	0.00
2	199	46	38.22	46	38.22	38.22	0.01	0.00

Continued on next page

Table 6.11 – *(Continued)*

Lead time	Lost sales penalty	Best base-stock		Limiting base-stock			Estimation Error (%)	Diff. from best base-stock (%)
		Level	Cost	Level	Estimated Cost	Real Cost		
3	1	7	4.24	7	4.24	4.24	0.11	0.00
3	4	18	11.13	18	11.13	11.13	0.03	0.00
3	9	26	16.87	26	16.87	16.87	0.01	0.00
3	19	33	22.73	33	22.73	22.73	-0.02	0.00
3	49	42	30.34	42	30.34	30.34	0.00	0.00
3	99	48	35.90	48	35.90	35.90	0.00	0.00
3	199	54	41.30	54	41.30	41.30	0.01	0.00
4	1	8	4.29	8	4.29	4.29	-0.04	0.00
4	4	21	11.44	21	11.44	11.44	-0.02	0.00
4	9	30	17.54	30	17.54	17.54	0.02	0.00
4	19	38	23.85	38	23.85	23.85	0.00	0.00
4	49	48	32.09	48	32.09	32.09	0.01	0.00
4	99	54	38.10	54	38.10	38.10	-0.01	0.00
4	199	61	43.91	61	43.91	43.91	-0.01	0.00

End of Table

Table 6.12 Performance of limiting base-stock policy for Poisson demand with varying mean and lead time of 2

penalty cost	mean demand	Best base-stock		Limiting base-stock			Estimation Error (%)	Diff. from best base-stock (%)
		Level	Cost	Level	Estimated Cost	Real Cost		
1	1	1	0.88	1	0.88	0.88	0.37	0.00
1	2	4	1.35	4	1.37	1.35	1.78	0.00
1	3	6	1.69	6	1.74	1.69	2.89	0.00
1	4	9	1.97	9	2.04	1.97	3.54	0.00
1	5	12	2.23	12	2.31	2.23	3.69	0.00
1	6	14	2.45	14	2.55	2.45	4.13	0.00
1	7	17	2.66	17	2.77	2.66	4.09	0.00
1	8	20	2.85	20	2.98	2.85	4.40	0.00
1	9	23	3.04	23	3.17	3.04	4.34	0.00
1	10	25	3.21	25	3.35	3.21	4.44	0.00
4	1	3	2.04	3	2.06	2.04	1.06	0.00
4	2	7	2.93	7	2.97	2.93	1.33	0.00
4	3	10	3.58	10	3.64	3.58	1.74	0.00
4	4	13	4.14	13	4.22	4.14	1.91	0.00

Continued on next page

penalty cost	mean demand	Best base-stock		Limiting base-stock			Estimation Error (%)	Diff. from best base-stock (%)
		Level	Cost	Level	Estimated Cost	Real Cost		
4	5	16	4.64	16	4.73	4.64	1.98	0.00
4	6	19	5.09	19	5.20	5.09	2.13	0.00
4	7	22	5.51	23	5.62	5.52	1.75	0.24
4	8	26	5.89	26	6.00	5.89	1.91	0.00
4	9	29	6.24	29	6.36	6.24	2.00	0.00
4	10	32	6.57	32	6.71	6.57	2.13	0.00
9	1	4	2.91	4	2.93	2.91	0.75	0.00
9	2	8	4.02	8	4.06	4.02	0.99	0.00
9	3	12	4.91	12	4.96	4.91	0.96	0.00
9	4	15	5.64	15	5.71	5.64	1.29	0.00
9	5	19	6.32	19	6.38	6.32	1.00	0.00
9	6	22	6.88	22	6.97	6.88	1.24	0.00
9	7	25	7.43	25	7.53	7.43	1.35	0.00
9	8	29	7.95	29	8.04	7.95	1.11	0.00
9	9	32	8.4	32	8.51	8.40	1.27	0.00
9	10	35	8.84	35	8.97	8.84	1.43	0.00
19	1	5	3.68	5	3.71	3.68	0.85	0.00
19	2	9	5.09	10	5.12	5.16	-0.68	1.30
19	3	13	6.12	13	6.17	6.12	0.89	0.00
19	4	17	7.01	17	7.06	7.01	0.76	0.00
19	5	21	7.84	21	7.89	7.84	0.68	0.00
19	6	24	8.52	24	8.59	8.52	0.84	0.00
19	7	28	9.21	28	9.28	9.21	0.71	0.00
19	8	31	9.79	31	9.87	9.79	0.85	0.00
19	9	34	10.38	35	10.48	10.46	0.27	0.74
19	10	38	10.91	38	11.00	10.91	0.85	0.00
49	1	7	4.63	7	4.64	4.63	0.28	0.00
49	2	11	6.26	11	6.28	6.26	0.40	0.00
49	3	15	7.56	15	7.59	7.56	0.42	0.00
49	4	19	8.65	19	8.69	8.65	0.44	0.00
49	5	23	9.63	23	9.67	9.63	0.38	0.00
49	6	26	10.5	26	10.57	10.50	0.63	0.00
49	7	30	11.26	30	11.32	11.26	0.56	0.00
49	8	34	12.02	34	12.08	12.02	0.49	0.00
49	9	37	12.7	37	12.78	12.70	0.61	0.00
49	10	41	13.36	41	13.43	13.36	0.49	0.00
99	1	7	5.24	7	5.26	5.24	0.41	0.00
99	2	12	7.11	12	7.13	7.11	0.32	0.00
99	3	16	8.56	16	8.60	8.56	0.42	0.00
99	4	20	9.78	20	9.82	9.78	0.42	0.00
99	5	24	10.84	24	10.89	10.84	0.42	0.00
99	6	28	11.81	28	11.85	11.81	0.33	0.00
99	7	32	12.71	32	12.75	12.71	0.35	0.00
99	8	35	13.55	35	13.61	13.55	0.43	0.00
99	9	39	14.29	39	14.35	14.29	0.39	0.00
99	10	43	15.04	43	15.09	15.04	0.33	0.00

Continued on next page

penalty cost	mean demand	Best base-stock Level	Best base-stock Cost	Limiting base-stock Level	Limiting base-stock Estimated Cost	Limiting base-stock Real Cost	Estimation Error (%)	Diff. from best base-stock (%)
199	1	8	5.82	8	5.83	5.82	0.18	0.00
199	2	13	7.92	13	7.94	7.92	0.23	0.00
199	3	17	9.5	17	9.53	9.50	0.31	0.00
199	4	21	10.86	21	10.90	10.86	0.34	0.00
199	5	25	12.03	25	12.07	12.03	0.37	0.00
199	6	29	13.07	29	13.12	13.07	0.35	0.00
199	7	33	14.02	33	14.07	14.02	0.33	0.00
199	8	37	14.92	37	14.96	14.92	0.27	0.00
199	9	41	15.79	41	15.82	15.79	0.21	0.00
199	10	44	16.6	44	16.66	16.60	0.35	0.00

End of Table

Table 6.13 Performance of limiting base-stock policy for Poisson demand with varying mean and lead time of 4

penalty cost	mean demand	Best base-stock Level	Best base-stock Cost	Limiting base-stock Level	Limiting base-stock Estimated Cost	Limiting base-stock Real Cost	Estimation Error (%)	Diff. from best base-stock (%)
1	1	2	0.91	2	0.92	0.91	0.65	0.00
1	2	6	1.42	5	1.45	1.42	1.86	0.06
1	3	10	1.79	9	1.84	1.79	2.73	0.11
1	4	14	2.09	14	2.17	2.09	3.65	0.00
1	5	18	2.37	18	2.46	2.37	3.62	0.00
1	6	22	2.61	22	2.72	2.61	4.08	0.00
1	7	27	2.83	27	2.96	2.83	4.58	0.00
1	8	31	3.04	31	3.18	3.04	4.40	0.00
1	9	36	3.24	36	3.39	3.24	4.54	0.00
1	10	40	3.43	40	3.58	3.43	4.52	0.00
4	1	5	2.24	5	2.26	2.24	0.89	0.00
4	2	10	3.23	10	3.28	3.23	1.58	0.00
4	3	15	3.99	15	4.06	3.99	1.82	0.00
4	4	20	4.62	20	4.71	4.62	2.01	0.00
4	5	25	5.19	25	5.29	5.19	1.88	0.00
4	6	30	5.69	30	5.81	5.69	2.02	0.00
4	7	35	6.13	35	6.29	6.13	2.48	0.00
4	8	40	6.59	40	6.73	6.59	2.07	0.00
4	9	45	6.98	45	7.15	6.98	2.38	0.00
4	10	50	7.39	50	7.54	7.39	2.04	0.00

Continued on next page

penalty cost	mean demand	Best base-stock		Limiting base-stock			Estimation Error (%)	Diff. from best base-stock (%)
		Level	Cost	Level	Estimated Cost	Real Cost		
9	1	6	3.23	6	3.26	3.23	0.98	0.00
9	2	12	4.57	12	4.63	4.57	1.37	0.00
9	3	17	5.62	17	5.71	5.62	1.61	0.00
9	4	23	6.48	23	6.57	6.48	1.51	0.00
9	5	28	7.24	28	7.36	7.24	1.55	0.00
9	6	34	7.96	34	8.06	7.96	1.33	0.00
9	7	39	8.52	39	8.70	8.52	2.04	0.00
9	8	44	9.17	44	9.30	9.17	1.48	0.00
9	9	49	9.70	50	9.87	9.72	1.63	0.15
9	10	55	10.28	55	10.40	10.28	1.17	0.00
19	1	8	4.21	8	4.26	4.21	1.27	0.00
19	2	14	5.88	14	5.94	5.88	1.11	0.00
19	3	20	7.15	20	7.26	7.15	1.55	0.00
19	4	25	8.24	25	8.35	8.24	1.41	0.00
19	5	31	9.17	31	9.30	9.17	1.45	0.00
19	6	37	10.06	37	10.18	10.06	1.22	0.00
19	7	42	10.76	42	10.97	10.76	1.97	0.00
19	8	48	11.59	48	11.72	11.59	1.11	0.00
19	9	53	12.23	53	12.41	12.23	1.53	0.00
19	10	58	12.98	59	13.08	12.98	0.80	0.02
49	1	9	5.29	9	5.42	5.29	2.33	0.00
49	2	16	7.43	16	7.54	7.43	1.37	0.00
49	3	22	8.90	22	9.12	8.90	2.54	0.00
49	4	28	10.34	28	10.47	10.34	1.28	0.00
49	5	34	11.44	34	11.66	11.44	1.86	0.00
49	6	40	12.52	40	12.73	12.52	1.68	0.00
49	7	46	13.40	46	13.72	13.40	2.45	0.00
49	8	51	14.46	51	14.64	14.46	1.27	0.00
49	9	57	15.19	57	15.48	15.19	1.95	0.00
49	10	63	16.20	63	16.30	16.20	0.65	0.00
99	1	10	6.02	10	6.25	6.02	3.75	0.00
99	2	17	8.45	17	8.62	8.45	2.05	0.00
99	3	24	10.03	24	10.47	10.03	4.33	0.00
99	4	30	11.79	30	11.95	11.79	1.38	0.00
99	5	36	12.93	36	13.28	12.93	2.64	0.00
99	6	42	14.07	42	14.48	14.07	2.89	0.00
99	7	48	15.04	48	15.59	15.04	3.65	0.00
99	8	54	16.40	54	16.63	16.40	1.39	0.00
99	9	59	17.15	60	17.61	17.16	2.63	0.06
99	10	65	18.35	65	18.51	18.35	0.88	0.00
199	1	11	6.59	11	7.03	6.59	6.62	0.00
199	2	18	9.31	18	9.65	9.31	3.62	0.00
199	3	25	10.74	25	11.63	10.74	8.26	0.00
199	4	31	13.08	31	13.34	13.08	1.99	0.00
199	5	38	14.20	38	14.80	14.20	4.22	0.00
199	6	44	15.33	44	16.11	15.33	5.10	0.00
199	7	50	16.35	50	17.32	16.35	5.97	0.00

Continued on next page

penalty cost	mean demand	Best base-stock Level	Best base-stock Cost	Limiting base-stock Level	Limiting base-stock Estimated Cost	Limiting base-stock Real Cost	Estimation Error (%)	Diff. from best base-stock (%)
199	8	56	18.10	56	18.46	18.10	1.99	0.00
199	9	62	18.70	62	19.53	18.70	4.43	0.00
199	10	68	20.23	68	20.55	20.23	1.56	0.00

End of Table

Table 6.14 Performance of limiting base-stock policy for Negative Binomial (r, q) demand and lead time of 2

penalty cost	Negative binomial parameter r	Negative binomial parameter q	Best base-stock level	Best base-stock cost	Limiting base-stock level	Limiting base-stock estimated cost	Limiting base-stock real cost	Estimation error (%)	Diff. from best base-stock (%)
1	1	0.1	11	7.27	11	7.27	7.27	0.02	0.00
1	1	0.2	5	3.40	5	3.40	3.40	0.13	0.00
1	1	0.3	2	2.09	2	2.09	2.09	0.02	0.00
1	1	0.4	1	1.41	1	1.41	1.41	-0.06	0.00
1	1	0.5	0	1.00	0	1.00	1.00	0.00	0.00
1	2	0.1	32	11.57	32	11.66	11.57	0.81	0.00
1	2	0.2	14	5.44	14	5.48	5.44	0.80	0.00
1	2	0.3	8	3.38	7	3.40	3.38	0.60	0.06
1	2	0.4	4	2.33	4	2.34	2.33	0.35	0.00
1	2	0.5	3	1.67	3	1.68	1.67	0.33	0.00
4	1	0.1	27	18.57	27	18.57	18.57	-0.02	0.00
4	1	0.2	12	8.73	12	8.73	8.73	0.02	0.00
4	1	0.3	7	5.42	7	5.42	5.42	0.01	0.00
4	1	0.4	4	3.76	4	3.76	3.76	-0.01	0.00
4	1	0.5	3	2.69	3	2.69	2.69	0.09	0.00
4	2	0.1	57	27.23	57	27.38	27.23	0.54	0.00
4	2	0.2	25	12.83	25	12.90	12.83	0.55	0.00
4	2	0.3	15	7.99	15	8.03	7.99	0.45	0.00
4	2	0.4	10	5.54	10	5.57	5.54	0.49	0.00
4	2	0.5	6	4.03	6	4.05	4.03	0.42	0.00
9	1	0.1	38	27.71	39	27.71	27.76	-0.18	0.17
9	1	0.2	18	13.06	18	13.06	13.06	0.00	0.00
9	1	0.3	10	8.13	10	8.13	8.13	0.06	0.00
9	1	0.4	7	5.62	7	5.62	5.62	0.05	0.00
9	1	0.5	5	4.10	5	4.10	4.10	0.00	0.00

Continued on next page

penalty cost	Negative binomial parameter r	q	Best base-stock Best base-stock level	cost	Limiting base-stock Limiting base-stock level	estimated cost	real cost	Estimation error (%)	Diff. from best base-stock (%)
9	2	0.1	73	39.17	73	39.32	39.17	0.37	0.00
9	2	0.2	33	18.46	33	18.53	18.46	0.37	0.00
9	2	0.3	19	11.52	20	11.56	11.52	0.37	0.00
9	2	0.4	13	7.98	13	8.01	7.98	0.40	0.00
9	2	0.5	9	5.83	9	5.85	5.83	0.26	0.00
19	1	0.1	49	36.92	49	36.92	36.92	-0.01	0.00
19	1	0.2	23	17.41	23	17.41	17.41	-0.01	0.00
19	1	0.3	14	10.86	14	10.86	10.86	0.01	0.00
19	1	0.4	9	7.52	9	7.52	7.52	-0.03	0.00
19	1	0.5	6	5.49	6	5.49	5.49	0.07	0.00
19	2	0.1	87	50.81	87	50.95	50.81	0.27	0.00
19	2	0.2	39	23.96	39	24.03	23.96	0.28	0.00
19	2	0.3	24	14.95	24	14.99	14.95	0.26	0.00
19	2	0.4	16	10.39	16	10.41	10.39	0.22	0.00
19	2	0.5	11	7.58	11	7.60	7.58	0.26	0.00
49	1	0.1	63	48.86	63	48.86	48.86	0.01	0.00
49	1	0.2	29	23.04	29	23.04	23.04	0.00	0.00
49	1	0.3	17	14.39	17	14.39	14.39	0.00	0.00
49	1	0.4	12	9.99	12	9.99	9.99	0.02	0.00
49	1	0.5	8	7.31	8	7.31	7.31	0.00	0.00
49	2	0.1	103	65.50	103	65.62	65.50	0.19	0.00
49	2	0.2	47	30.89	47	30.95	30.89	0.19	0.00
49	2	0.3	28	19.30	29	19.33	19.33	0.00	0.15
49	2	0.4	19	13.40	19	13.42	13.40	0.17	0.00
49	2	0.5	13	9.83	13	9.84	9.83	0.15	0.00
99	1	0.1	72	57.63	72	57.63	57.63	0.00	0.00
99	1	0.2	33	27.18	33	27.18	27.18	0.00	0.00
99	1	0.3	20	19.96	20	16.96	19.96	-15.01	0.00
99	1	0.4	14	11.81	14	11.81	11.81	0.00	0.00
99	1	0.5	10	8.65	10	8.65	8.65	-0.01	0.00
99	2	0.1	115	76.08	115	76.19	76.08	0.14	0.00
99	2	0.2	53	35.89	53	35.94	35.89	0.13	0.00
99	2	0.3	32	22.40	32	22.43	22.40	0.13	0.00
99	2	0.4	21	15.61	21	15.63	15.61	0.13	0.00
99	2	0.5	15	11.41	15	11.42	11.41	0.11	0.00
199	1	0.1	81	66.18	81	66.18	66.18	0.01	0.00
199	1	0.2	37	31.23	37	31.23	31.23	0.00	0.00
199	1	0.3	23	19.49	23	19.49	19.49	-0.01	0.00
199	1	0.4	16	13.60	16	13.60	13.60	-0.03	0.00
199	1	0.5	11	9.93	11	9.93	9.93	-0.03	0.00
199	2	0.1	126	86.26	126	86.35	86.26	0.10	0.00
199	2	0.2	58	40.69	58	40.73	40.69	0.10	0.00
199	2	0.3	35	25.40	35	25.43	25.40	0.11	0.00
199	2	0.4	24	17.69	24	17.70	17.69	0.08	0.00
199	2	0.5	17	12.96	17	12.97	12.96	0.10	0.00

Continued on next page

	Negative binomial parameter		Best base-stock Best base-stock		Limiting base-stock Limiting base-stock				
penalty cost	r	q	level	cost	level	estimated cost	real cost	Estimation error (%)	Diff. from best base-stock (%)

End of Table

Table 6.15 Performance of limiting base-stock policy for Negative Binomial (r, q) demand and lead time of 4

	Negative binomial parameter		Best base-stock		Limiting base-stock				
penalty cost	r	q	level	cost	level	estimated cost	real cost	Estimation error (%)	Diff. from best base-stock (%)
1	1	0.1	16	7.46	16	7.46	7.46	0.06	0.00
1	1	0.2	6	3.49	6	3.49	3.49	0.04	0.00
1	1	0.3	3	2.13	3	2.13	2.13	0.01	0.00
1	1	0.4	1	1.44	1	1.44	1.44	0.04	0.00
1	1	0.5	1	1.00	1	1.00	1.00	0.06	0.00
1	2	0.1	48	12.02	48	12.13	12.02	0.86	0.00
1	2	0.2	20	5.65	20	5.70	5.65	0.87	0.00
1	2	0.3	11	3.50	11	3.53	3.50	0.66	0.00
1	2	0.4	6	2.40	6	2.42	2.40	0.64	0.00
1	2	0.5	4	1.72	4	1.73	1.72	0.39	0.00
4	1	0.1	39	19.80	38	19.84	19.80	0.23	0.00
4	1	0.2	17	9.31	17	9.33	9.31	0.17	0.00
4	1	0.3	10	5.79	10	5.79	5.79	0.10	0.00
4	1	0.4	6	3.97	6	3.98	3.97	0.17	0.00
4	1	0.5	4	2.86	4	2.87	2.86	0.32	0.00
4	2	0.1	83	29.54	84	29.71	29.54	0.58	0.00
4	2	0.2	37	13.89	37	13.99	13.89	0.72	0.00
4	2	0.3	21	8.66	21	8.71	8.66	0.57	0.00
4	2	0.4	14	5.99	14	6.02	5.99	0.60	0.00
4	2	0.5	9	4.35	9	4.36	4.35	0.44	0.00
9	1	0.1	54	30.28	54	30.41	30.28	0.43	0.00
9	1	0.2	24	14.30	24	14.32	14.30	0.11	0.00
9	1	0.3	14	8.91	14	8.91	8.91	0.00	0.00
9	1	0.4	9	6.13	9	6.16	6.13	0.44	0.00
9	1	0.5	6	4.45	6	4.47	4.45	0.45	0.00
9	2	0.1	105	43.68	105	43.85	43.68	0.39	0.00
9	2	0.2	47	20.56	47	20.66	20.56	0.51	0.00

Continued on next page

penalty cost	Negative binomial parameter r	q	Best base-stock level	Best base-stock cost	Limiting base-stock level	Limiting base-stock estimated cost	Limiting base-stock real cost	Estimation error (%)	Diff. from best base-stock (%)
9	2	0.3	28	12.82	28	12.88	12.82	0.42	0.00
9	2	0.4	18	8.88	18	8.92	8.88	0.51	0.00
9	2	0.5	12	6.48	12	6.50	6.48	0.24	0.00
19	1	0.1	68	40.98	67	41.32	40.98	0.83	0.02
19	1	0.2	30	19.47	31	19.48	19.47	0.03	0.02
19	1	0.3	18	12.13	18	12.13	12.13	0.01	0.00
19	1	0.4	12	8.34	12	8.41	8.34	0.84	0.00
19	1	0.5	8	6.08	8	6.13	6.08	0.78	0.00
19	2	0.1	123	57.87	124	58.00	57.87	0.23	0.01
19	2	0.2	56	27.25	56	27.34	27.25	0.33	0.00
19	2	0.3	33	16.99	33	17.05	16.99	0.36	0.00
19	2	0.4	22	11.76	22	11.83	11.76	0.63	0.00
19	2	0.5	15	8.62	15	8.63	8.62	0.11	0.00
49	1	0.1	84	54.60	84	55.57	54.60	1.78	0.00
49	1	0.2	38	26.31	38	26.20	26.31	-0.40	0.00
49	1	0.3	23	16.42	23	16.34	16.42	-0.48	0.00
49	1	0.4	15	11.17	15	11.37	11.17	1.74	0.00
49	1	0.5	11	8.14	11	8.29	8.14	1.85	0.00
49	2	0.1	145	76.19	145	76.00	76.19	-0.24	0.00
49	2	0.2	66	35.82	66	35.83	35.82	0.05	0.00
49	2	0.3	40	22.32	40	22.36	22.32	0.19	0.00
49	2	0.4	26	15.38	26	15.54	15.38	1.03	0.00
49	2	0.5	18	11.37	18	11.36	11.37	-0.08	0.00
99	1	0.1	96	64.13	96	65.98	64.13	2.87	0.00
99	1	0.2	44	31.51	44	31.11	31.51	-1.29	0.00
99	1	0.3	26	19.70	27	19.43	19.71	-1.45	0.08
99	1	0.4	18	13.10	18	13.49	13.10	2.98	0.00
99	1	0.5	13	9.61	13	9.90	9.61	3.00	0.00
99	2	0.1	160	89.65	160	88.88	89.65	-0.86	0.00
99	2	0.2	73	42.05	73	41.91	42.05	-0.32	0.00
99	2	0.3	44	26.16	44	26.16	26.16	0.00	0.00
99	2	0.4	29	17.88	29	18.19	17.88	1.75	0.00
99	2	0.5	20	13.36	20	13.34	13.36	-0.16	0.00
199	1	0.1	106	72.71	107	76.02	72.75	4.48	0.06
199	1	0.2	49	36.89	49	35.84	36.89	-2.85	0.00
199	1	0.3	30	23.09	30	22.38	23.09	-3.11	0.00
199	1	0.4	20	14.84	20	15.55	14.84	4.79	0.00
199	1	0.5	14	10.84	14	11.39	10.84	5.05	0.00
199	2	0.1	174	102.79	174	101.15	102.79	-1.59	0.00
199	2	0.2	79	47.99	79	47.71	47.99	-0.58	0.00
199	2	0.3	48	29.86	48	29.78	29.86	-0.28	0.00
199	2	0.4	32	20.11	32	20.71	20.11	2.95	0.00
199	2	0.5	23	15.22	23	15.20	15.22	-0.10	0.00

End of Table

Table 6.16 Performance of limiting base-stock policy for discretized generalized Pareto demand with $k = 0.1$ and $\sigma = 5$

Lead time	Lost sales penalty	Best base-stock		Limiting base-stock			Estimation Error (%)	Diff. from best base-stock (%)
		Level	Cost	Level	Estimated Cost	Real Cost		
1	1	5	4.28	5	4.28	4.28	-0.13	0.00
1	4	13	10.92	13	10.90	10.92	-0.23	0.00
1	9	18	16.45	18	16.44	16.45	-0.06	0.00
1	19	24	22.41	24	22.29	22.41	-0.53	0.00
1	49	32	30.45	32	30.46	30.45	0.01	0.00
1	99	38	37.70	38	36.97	37.70	-1.95	0.00
1	199	45	43.31	45	43.77	43.31	1.06	0.00
2	1	7	4.40	7	4.39	4.40	-0.21	0.00
2	4	16	11.56	16	11.54	11.56	-0.09	0.00
2	9	23	17.77	23	17.68	17.77	-0.49	0.00
2	19	30	24.35	30	24.19	24.35	-0.66	0.00
2	49	40	33.98	39	33.19	33.98	-2.32	0.01
2	99	46	40.28	46	40.24	40.28	-0.12	0.00
2	199	54	46.72	53	47.50	46.75	1.61	0.06
3	1	9	4.47	9	4.47	4.47	-0.07	0.00
3	4	20	12.01	20	11.97	12.01	-0.31	0.00
3	9	28	18.62	28	18.56	18.62	-0.34	0.00
3	19	36	25.63	36	25.62	25.63	-0.04	0.00
3	49	46	35.33	46	35.34	35.33	0.02	0.00
3	99	54	42.39	54	42.88	42.39	1.14	0.00
3	199	61	52.27	61	50.57	52.27	-3.25	0.00
4	1	10	4.51	10	4.51	4.51	-0.11	0.00
4	4	23	12.30	23	12.27	12.30	-0.28	0.00
4	9	32	19.30	32	19.23	19.30	-0.34	0.00
4	19	41	26.87	41	26.76	26.87	-0.41	0.00
4	49	52	37.39	52	37.13	37.39	-0.70	0.00
4	99	61	45.58	61	45.12	45.58	-1.02	0.00
4	199	69	53.57	69	53.21	53.57	-0.68	0.00

End of Table

Table 6.17 Performance of limiting base-stock policy for discretized generalized Pareto demand with $k = 0.4$ and $\sigma = 10$

Lead time	Lost sales penalty	Best base-stock		Limiting base-stock			Estimation Error (%)	Diff. from best base-stock (%)
		Level	Cost	Level	Estimated Cost	Real Cost		
1	1	12	13.89	12	13.87	13.89	-0.14	0.00
1	4	32	40.52	32	40.29	40.52	-0.56	0.00
1	9	52	68.20	52	67.72	68.20	-0.70	0.00
1	19	76	103.78	76	103.23	103.78	-0.53	0.00
1	49	118	167.22	117	166.50	167.22	-0.43	0.00
1	99	158	235.59	158	230.94	235.59	-1.97	0.00
1	199	211	314.23	210	314.62	314.24	0.12	0.00
2	1	15	14.16	15	14.12	14.16	-0.26	0.00
2	4	40	42.08	41	41.80	42.08	-0.68	0.01
2	9	64	71.29	64	70.81	71.29	-0.68	0.00
2	19	92	108.55	91	108.15	108.56	-0.38	0.01
2	49	137	174.83	137	173.71	174.83	-0.64	0.00
2	99	181	240.86	180	239.53	240.87	-0.56	0.01
2	199	237	320.30	234	324.18	320.34	1.20	0.01
3	1	19	14.30	19	14.26	14.30	-0.28	0.00
3	4	48	43.06	49	42.79	43.06	-0.62	0.01
3	9	75	73.57	75	73.04	73.57	-0.72	0.00
3	19	107	112.60	106	111.94	112.60	-0.58	0.00
3	49	155	182.81	155	179.62	182.81	-1.75	0.00
3	99	203	247.76	201	246.79	247.80	-0.41	0.01
3	199	258	337.44	257	332.43	337.45	-1.49	0.00
4	1	22	14.38	22	14.35	14.38	-0.23	0.00
4	4	56	43.84	56	43.51	43.84	-0.76	0.00
4	9	86	75.18	86	74.76	75.18	-0.56	0.00
4	19	120	116.45	120	115.02	116.45	-1.23	0.00
4	49	175	186.24	173	184.66	186.26	-0.86	0.01
4	99	223	253.74	221	253.14	253.77	-0.25	0.01
4	199	284	339.33	280	339.79	339.40	0.12	0.02

End of Table

Chapter 7

Conclusion

"O be wise, what can I say more?"

Jacob, the brother of Nephi

In this thesis, we studied maintenance spare parts planning and control. We provided a general framework for maintenance spare parts planning and control in chapter 2. In chapters 3-5, we paid particular attention to the way maintenance strategies affect the demand process for rotables and repairables. For rotables with a usage based maintenance strategy, this led to a deterministic planning problem as studied in chapter 3. Chapters 4 and 5 studied planning for repairables with either a breakdown corrective or condition based maintenance strategy. We studied mechanisms to expedite repairs as a way of providing lead time flexibility. Finally, chapter 6 considered a model for consumable inventory with emergency procedures in case of stock-outs.

7.1. Research objectives revisited

In the introduction of this thesis we stated 9 research objectives. Now we revisit each of these research objectives and summarize what we have learned.

7.1.1 Framework

Research objective 1 Develop a framework for the planning and control of a spare part supply chain in organizations that own and maintain equipment. This framework

should outline all relevant decisions that are made in such a supply chain and explain how they relate to each other.

In §2, we developed a framework that identifies eight main processes involved with controlling a spare part supply chain. We have also classified these decisions as either strategic, tactical of operational and explained how they relate to each other.

7.1.2 Rotables, usage based maintenance and efficient utilization of resources

Research objective 2 Develop a planning algorithm that makes efficient use of the resources needed for rotable replacement and overhaul. This algorithm should exploit the fact that demand for rotables is predictable.

In chapter 3, we developed a MIP formulation of the planning problem that turned out the be strongly $\mathcal{NP}$-hard. In spite of this, we provided computational evidence that the LP relaxation of our formulation gives sufficiently accurate results to guide decision making for instances of real life size.

The modeling phase also identified a flaw in the typical way the objective function for such problems is set up. Most approaches focus on maximizing the time a rotable spends in the field before it is overhauled. A better approach is to minimize the total number of times a rotable is overhauled during its entire lifetime. These two objectives are not equivalent because a rotable has finite lifetime. This insight also led us to consider an additional degree of freedom in the planning of rotable replacement and overhaul: It is possible to replace rotables earlier than the usage based maintenance strategy requires without increasing the total number of replacements and overhauls during the entire lifetime of a rotable. This extra planning flexibility can be used to efficiently utilize the finite resources needed for replacement and overhaul.

Research objective 3 Investigate the value of using the predictability of demand for rotables in making efficient use of resources.

A case study in chapter 3 based on NedTrain data identified labor costs as the dominant cost factor. The focus on performing replacement and overhaul as late as possible naturally leads to the need to employ a large workforce to deal with peaks in overhaul requirements. Our model not only uses that demand for rotables is predictable, but also smooths the demand for rotables. This smoothing significantly decreases the size of the required workforce, and leads to cost savings of around 4% in the case study. There is an essential trade-off between smoothing the workload in the

overhaul workshop and performing overhaul as late as possible. For the case study at NedTrain, we found that it is more appropriate to focus on smoothing the workload.

7.1.3 Repairables, condition based maintenance, and repair lead time flexibility

Research objective 4 Develop a model of repair lead time flexibility and non-stationary demand due to condition based maintenance for a single-item and investigate how information regarding demand non-stationarity from condition based maintenance can be used to leverage repair lead time flexibility.

In chapter 4, we developed a model in which lead time flexibility is modeled through an expediting option. Non-stationarity of demand as a consequence of condition based maintenance was modeled by a Markov modulated Poisson process.

We studied the optimal expediting policy and found that it has the following form: Keep the number of parts on-hand and arriving to inventory shortly above some threshold level that depends on the pipeline of outstanding repair orders and the distribution of demand for the near future. Unfortunately, the optimal policy turned out to be computationally intractable, so we also investigated a heuristic expediting policy that aggregates information about the pipeline of repair orders. This policy was shown to perform well in a computational study.

Research objective 5 Develop a model that can assess the interplay between repairable inventory and lead time flexibility in buffering demand uncertainty and non-stationarity.

In chapter 4, we also studied the joint optimization of repairable stock levels and expediting policy. We showed formally that keeping more stock leads to less expediting and vice versa in optimal solutions.

Research objective 6 Investigate the value of explicitly modeling lead time flexibility and demand information arising from condition based maintenance.

In a numerical study for the single item model in chapter 4, we compared the optimal stocking and expediting decision with decisions based on optimization models that do not explicitly model lead time flexibility or demand information arising from condition based maintenance. We found that this leads to average optimality gaps of more than 11% for single-item problems in a wide test bed. Furthermore, maximum optimality gaps ranged all the way up to 64% in this test bed.

For multi-item problems as considered in chapter 5, the possibility to expedite could be used to reduce the investment in stock by 29% on average compared to the situation where the mean lead time was identical, but expediting is not possible.

Research objective 7 Develop a tractable multi-item optimization algorithm that supports the initial supply decision and incorporates lead time flexibility, non-stationary demand arising from condition based maintenance and performance objectives on fleet level.

Chapter 5 describes an algorithm that supports stocking decisions (including initial supply) and incorporates both lead time flexibility (through expediting) and non-stationary demand modeled by a Markov modulated Poisson process. The average optimality gap of solutions found by this algorithm was 0.62% and 8.85% at most in a computational study. The computational time was just over a minute on average and always below 10 minutes for large instances.

7.1.4 Consumables, emergency procedures

Research objective 8 Develop an algorithm for the optimization of the base-stock level in a periodic review lost sales inventory system that is fast and provides accurate estimates of performance measures.

In chapter 6, we developed a fast algorithm to approximately evaluate and optimize base-stock levels for the periodic review lost sales inventory systems. In numerical work, the approximation was shown to be exceptionally accurate. This can be partly explained by asymptotic results also given in chapter 6.

Bibliography

J. Abate and W. Whitt. Numerical inversion of probability generating functions. *Operations Research Letters*, 12:245–251, 1992.

AberdeenGroup. Service parts management, unlocking value and profits in the service chain. *AberdeenGroup, Boston*, 2003.

I.J.B.F. Adan, A. Sleptchenko, and G.J. Van Houtum. Reducing costs of spare parts supply systems via static priorities. *Asia-Pacific Journal of Operatonal Research*, 26(4):559–585, 2009.

Y. Akçay and S.H. Xu. Joint inventory replenishment and component allocation optimization in an assemble-to-order system. *Management Science*, 50(1):99–116, 2004.

D. Aldous and L. Shepp. The least variable phase type distribution is Erlang. *Communications in Statistics: Stochastic Models*, 3(3):467–473, 1987.

P. Alfredsson. Optimization of multi-echelon repairable item inventory systems with simulataneous location of repair facilities. *European Journal of Operational Research*, 99:584–595, 1997.

P. Alfredsson and J. Verrijdt. Modeling emergency supply flexibility in a two-echelon inventory system. *Management Science*, 45(10):1416–1431, 1999.

N. Altay and L.A. Litteral, editors. *Service Parts Management: Demand Forecasting and Inventory Control*. Springer, 2011.

E. Altman and G. Koole. On submodular value functions and complex dynamic programming. *Stochastic Models*, 14(5):1051–1072, 1998.

E.M. Alvarez, M.C. van der Heijden, and W.H.M. Zijm. The selective use of emergency shipments for service-contract differentiation. *International Journal of Production Economics*, 143:518–526, 2013a.

E.M. Alvarez, M.C. van der Heijden, and W.H.M. Zijm. Service differentiation in

spare parts supply through dedicated stocks. *Annals of Operations Research*, in press, 2013b. URL http://dx.doi.org/10.1007/s10479-013-1362-z.

J. Arts and S.D. Flapper. Aggregate overhaul and supply chain planning for rotables. *Annals of Operations Research*, forthcoming, 2013. URL http://dx.doi.org/10.1007/s10479-013-1426-0.

J. Arts, M. Van Vuuren, and G.P. Kiesmüller. Efficient optimization of the dual index policy using Markov chains. *IIE Transactions*, 43(8):604–620, 2011.

Y. Asiedu and P. Gu. Product life cycle cost analysis: state of the art review. *Internationa Journal of Production Research*, 36(4):883–908, 1998.

L.L. Barros and M. Riley. A combinatorial approach to level of repair analysis. *European Journal of Operational Research*, 129(2):242–251, 2001.

R.J.I. Basten and G.J. Van Houtum. System-oriented inventory models for spare parts. *Working Paper, Twente University*, 2013.

R.J.I. Basten, J.M.J. Schutten, and M.C. van der Heijden. An efficient model formulation for level of repair analysis. *Annals of Operations Research*, 172(1): 119–142, 2009.

R.J.I. Basten, M.C. van der Heijden, and J.M.J. Schutten. A minimum cost flow model for level of repair analysis. *International Journal of Production Economics*, 133(1):233–242, 2011.

J.W.M. Bertrand, J.C. Wortman, and J. Wijngaard. *Production control : a structural and design oriented approach*. Elsevier, 1990.

M. Bijvank and S.G. Johansen. Periodic review lost-sales inventory models with compound poisson demand and constant lead times of any length. *European Journal of Operational Research*, 220:106–114, 2012.

P.J. Billington, J.O. McClain, and L.J. Thomas. Mathematical programming approaches to capcity-contrained MRP systems: Review, formualtion and problem reduction. *Management Science*, 29(10):1126–1141, 1983.

G.R. Bitran and A.C. Hax. On the design of hierarchical production planning systems. *Decision Sciences*, 8:28–55, 1977.

G.R. Bitran, E.A. Haas, and A.C. Hax. Hierarchical production planning: a single stage system. *Operations Research*, 29(4):717–743, 1981.

G.R. Bitran, E.A. Haas, and A.C. Hax. Hierarchical production planning: a two stage system. *Operations Research*, 30(2):232–251, 1982.

S. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning large call centers.

Operations Research, 52(1):17–34, 2004.

G.E.P. Box and G.M. Jenkins. *Time series analysis, forecasting and control*. San Francisc0: Holden-Day, 1970.

J.R. Bradley and H.H. Guerrero. Lifetime buy decisions with multiple obsolete parts. *Production and Operations Management*, 18(1):114–126, 2009.

R. Brooks and A. Geoffrion. Finding Everett's Lagrange, multipliers by linear programming. *Operations Research*, 14(6):1149–1153, 1966.

K.E. Caggiano, J.A. Muckstadt, and J.A. Rappold. Integrated real-time capacity and inventory allocation for reparable service parts in a two-echelon supply system. *Manufacturing & Service Operations Management*, 8(3):292–319, 2006.

M.J. Carillo. Generalizations of Palm's theorem and Dyna-METRIC's demand and pipeline variability. *RAND report*, R-3698-AF, 1989.

S.R. Chakravarthy and A. Agarwal. Analysis of a machine repair problem with an unreliable server and phase type repairs and services. *Naval Research Logistics*, 50 (5):462–480, 2003.

M. Charest and J.A. Ferland. Preventive maintenance scheduling of power generating units. *Annals of Operations Research*, 41:185–206, 1993.

C. Chatfield. *The analysis of time series: an introduction*. Chapman & Hall/CRC, 2004.

C.H. Chen, S. Yan, and M. Chen. Short-term manpower planning for MRT carriage maintenance under mixed deterministic and stocastic demands. *Annals of Operations Research*, 181:67–88, 2010.

P.Y. Cho. Optimal scheduling of fighter aircraft maintenance. Master's thesis, Sloan School of Management, Massachusetts Institute of Technology, 2011.

J. Coetzee. *Maintenance*. Trafford publishing, 1997.

Supply Chain Council. Supply Chain Operations Reference (SCOR) model, 2010. URL `http://supply-chain.org/f/SCOR-Overview-Web.pdf`.

J.D. Croston. Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, 23(3):289–303, 1972.

G.B. Dantzig and P. Wolfe. Decomposition principle for linear programs. *Operations Research*, 8:101–111, 1960.

T.G. De Kok and J.C. Fransoo. Planning supply chain operations: Definition and comparison of planning concepts. In A.G. de Kok and S.C. Graves, editors, *Supply Chain Management: Design, Coordination and Operation*, pages 597–675. Elsevier,

2003.

Deloitte. The service revolution in global manufacturing industries. *Deloitte Research*, 2006.

M.A.H. Dempster, M.L. Fisher, Jansen L., Lageweg B.J., Lenstra J.K., and Rinnooy Kan A.H.G. Analytical evaluation of hierarchical planning systems. *Operations Research*, 29(4):707–716, 1981.

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

U. Dinesh Kumar, J. Crocker, J. Knezevic, and M. El-Haram. *Reliability Maintenance and Logistic Support - A Life Cycle Approach.* Kluwer academic publishers, 2000.

R. Downs, R. Metters, and J. Semple. Manageging inventory with multiple products, lags in delivery, resource constraints and lost sales: A mathematical programming approach. *Management Science*, 47(3):464–479, 2001.

M. Driessen. Logistiek Centrum Woensdrecht - Case Studie. available upon request from the author, October 2011.

M. Driessen. KLM Engineering and Maintenance - Case Study. available upon request from the author, February 2012a.

M. Driessen. Defensie Bedrijf Grondgebonden Systemen - Case Studie. available upon request by the author, October 2012b.

M. Driessen. Marinebedrijf - Case Studie. available upon request from the author, October 2012c.

M. Driessen. Europe Container Terminals - Case Studie. available upon request from the author, February 2013.

M. Driessen and J. Arts. NedTrain - Case Study. Eindhoven University of Technology, available upon request from the authors, May 2011.

M.A. Driessen, J.J. Arts, G.J. Van Houtum, W.D. Rustenburg, and B. Huisman. Maintenance spare part planning and control: A framework for control and agenda for future research. *Beta Working paper 325, Eindhoven University of Technology*, 2010.

B.P. Dzielinski, C.T. Baker, and A.S. Manne. Simulation tests of lot size programming. *Management Science*, 9(2):229–258, 1963.

C.E. Ebeling. *Introduction to reliability and maintainablity engineering.* McGraw-Hill, 2nd edition, 2001.

A.H. Elwany and N.Z. Gebraeel. Sensor-driven prognostic models for equipment replacement and spare parts inventory. *IIE Transactions*, 40(7):629–639, 2008.

Ş.S. Erengüç, N.C. Simpson, and A.J. Vakharia. Integrated production/distribution planning in supply chains: An invited review. *European Journal of Operational Research*, 115(2):219–236, 1999.

H. Everett. Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Operations Research*, 11(3):399–417, 1963.

Q. Feng, S.P. Sethi, H. Yan, and H. Zhang. Are base-stock policies optimal in inventory problems with multiple delivery modes? *Operations Research*, 54(4): 801–807, 2006.

W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.

M. Fischetti, F. Glover, and A. Lodi. The feasibility pump. *Mathematical Programming*, 10(1):91–104, 2005.

M.L. Fisher. The Lagrangian relaxtion method for solving integer programming problems. *Management Science*, 27(1):1–18, 1981.

S. Foss, K. Korshunov, and S. Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer Series in Operations Research and Financial Engineering. Springer, 2011.

J.C. Fransoo and V.C.S. Wiers. Action variety of planners: Cognitive load and requisite variety. *Journal of Operations Management*, 24:813–821, 2006.

J.C. Fransoo and V.C.S. Wiers. An empirical investigation of the neglect of mrp information by production planners. *Production Planning & Control*, 19(8):781–787, 2008.

Y. Fukuda. Optimal policies for the inventory problem with negotiable leadtime. *Management Science*, 10(4):690–708, 1964.

M.R. Garey and D.S. Johnson. *Computers and intractibility: A guide to the theory of NP-completeness*. W.H. Freeman and company, 1979.

G.M. Gaukler, Ö. Özer, and W.H. Hausman. Order progress information: Improved dynamic emergency ordering policies. *Propduction and Operations Management*, 17(6):599–613, 2008.

H.A. Ghoneim and S. Stidham Jr. Control of arrivals to two queues in series. *European Journal of Operational Research*, 21:399–409, 1985.

D.A. Goldberg, D.A. Katz-Rogozhnikov, Y. Lu, M. Sharma, and M.S. Squillante.

Asymptotic optimality of constant-order policies for lost sales inventory models with large lead times. *arXiv:1211.4063*, pages 1–17, 2012.

S.C. Graves. A multi-echelon inventory model for a repairable item with one-for-one replenishments. *Management Science*, 31(10):1247–1256, 1985.

F. Gross and J.F. Ince. Spares provisioning for repairable items: Cyclic queues in light traffic. *IIE Transactions*, 10(3):307–314, 1978.

V.D.R. Guide Jr. and R. Srivastava. Repairable inventory theory: Models and applications. *European Journal of Operational Research*, 102:1–20, 1997.

V.D.R. Guide Jr, R. Srivastava, and M.E. Kraus. Priority scheduling policies for repair shops. *International Journal of Production Research*, 38(4):929–950, 2000.

Y. Gupta and W.S. Chow. Twenty-five years of life cycle costing - theory and applications: A survey. *International Journal of Quality and Reliability Management*, 2(3):51–76, 1985.

W.H. Hausman and G.D. Scudder. Priority scheduling rules for repairable inventory systems. *Management Science*, 28(11):1215–1232, 1982.

A.C. Hax. Aggregate production planning. In J.J. Moder and S.E. Elmaghraby, editors, *Handbook of operations research models and applications*, pages 53–69. Van Nostrand-Reinhold Co, New York, 1978.

A.C. Hax and H.C. Meal. Hierarchical integration of production planning and scheduling. In M.A. Geisler, editor, *Studies in Management Sciences Vol. 1 Logistics*, pages 53–69. North-Holland-American Elsevier, 1975.

K.L. Head, P.B. Mirchandani, and D. Sheppard. Hierarchical framework for real-time traffic control. *Transportation research record*, 1360:82–88, 1992.

H. Heffes and D.M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 4(6):856–868, 1986.

A. Heng, S. Zhang, A.C.C. Tan, and J. Mathew. Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, 23:724739, 2009.

R.J. Hillestad. Dyna-METRIC: Dynamic Multi-Echelon Technique for Recoverable Item Control. *RAND report*, R-2785-AF, 1982.

W.J. Hopp and M.L. Spearman. *Factory physics*. McGraw-Hill, 2001.

D.A. Hounshell. *From the American system to mass production 1800-1932*. The Johns Hopkins University Press, 1984.

Z.S. Hua, B. Zhang, J. Yang, and D.S. Tan. A new approach of forecasting intermittent demand for spare parts inventories in the process industries. *Journal of the Operational Research Society*, 58:52–61, 2007.

W.T. Huh, G. Janakiraman, J.A. Muckstadt, and P. Rusmevichientong. An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research*, 34(2):397–416, 2009a.

W.T. Huh, G. Janakiraman, J.A. Muckstadt, and P. Rusmevichientong. Asymptotic optimality of order-up-to policies in lost sales inventory systems. *Management Science*, 55(3):404–420, 2009b.

J. Huiskonen. Maintenance spare parts logistics: Special characteristics and strategic choices. *International Journal of Production Economics*, 71:125–133, 2001.

D.L. Iglehart. Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability*, 2:429–441, 1965.

M.A. Ilgin and S.M. Gupta. Environmentally conscious manufacturing and product recovery (ECMPRO): A review of the state of the art. *Journal of Environmental Management*, 91:563–591, 2010.

K.E. Isaacson and P.M. Boren. Dyna-METRIC version 6: An advanced capability assessment model. *RAND report*, R-4214-AF, 1993.

D. Ivanov. An adaptive framework for aligning (re)planning decisions on supply chain strategy, design, tactics and operations. *International Journal of Production Research*, 48(13):3999–4017, 2010.

G. Janakiraman and R. Roundy. Lost-sales problems with stochastic lead times: convexity results for base-stock policies. *Operations Research*, 52(5):795–803, 2004.

G. Janakiraman, S. Seshadri, and G. Shantikumar. A comparison of the optimal costs of two canonical inventory systems. *Operations Research*, 55:866–875, 2007.

J.B. Jasper. Quick repsonse solutions: Fedex critical inventory logistics revitalized. *FedEx white paper*, 2006.

S.G. Johansen. Pure and modified base-stock policies for the lost sales inventory system with negligible set-up costs and constant lead times. *International Journal of Production Economics*, 71:391–399, 2001.

S.J. Joo. Scheduling preventive maintenance for modular designed components: A dynamic approach. *European Journal of Operational Research*, 192:512–520, 2009.

S. Karlin and H. Scarf. Inventory models of the Arrow-Harris-Marchak type with time lag. In K. Arrow, S. Karlin, and H. Scarf, editors, *Studies in the Mathematical Theory of Inventory and Production*. Stanford university press, Stanford, CA., 1958.

W.J. Kennedy, J.D. Patterson, and L.D. Fredendall. An overview of recent literature on spare parts inventories. *International Journal of Production Economics*, 76: 201–215, 2002.

T. Koch, T. Achterberg, E. Andersen, O. Bastert, T. Berthold, R.E. Bixby, E. Danna, G. Gamrath, A.M. Gleixner, S. Heinz, A. Lodi, H. Mittelmann, T. Ralphs, D. Salvagnin, D.E. Steffy, and K. Wolter. MIPLIB 2010. *Mathematical Programming Computation*, 3:103–163, 2011.

G. Koole. Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems*, 30:323–339, 1998.

G. Koole. Convexity in tandem queues. *Probability in the Engineering and Informational Sciences*, 18:13–31, 2004.

G. Koole. Monotonicity in Markov reward and decision chains: Theory and applications. *Foundations and Trends in Stochastic Systems*, 1(1):1–76, 2006.

A.A. Kranenburg and G.J. Van Houtum. Effect of commonality on spare part provisioning costs for capital goods. *International Journal of Production Economics*, 108:221–227, 2007.

A.A. Kranenburg and G.J. Van Houtum. Service differentiation in spare parts inventory management. *Journal of the Operational Research Society*, 59:946–955, 2008.

A.A. Kranenburg and G.J. Van Houtum. A new partial pooling structure for spare parts networks. *European Journal of Operational research*, 199(3):908–921, 2009.

V.G. Kulkarni. *Modeling, Analysis, Design, and Control of Stochastic Systems*. Springer, 1999.

H.C. Lau and H. Song. Multi-echelon repairable item inventory system with limited repair capcity under non-stationary demands. *International Journal of Inventory Research*, 1(1):67–92, 2008.

H.L. Lee. A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Science*, 33(10):1302–1316, 1987.

R. Levi, G. Janakiraman, and M. Nagaraja. A 2-approximation algorithm for stochastic inventory models with lost sales. *Mathematics of Operations Research*, 33(2):351–374, 2008.

Y. Lu, J.S. Song, and Y. Zhao. No-holdback allocation rules for continuous-time assemble-to-order systems. *Operations Research*, 58(3):691–705, 2010.

M.E. Lübbecke and J. Desrosiers. Selected topics in column generation. *Operations Research*, 53(6):1007–1023, 2005.

H.C. Meal. Putting production decisions where they belong. *Harvard Business Review*, 62:102–111, 1984.

K.S. Meier-Hellstern. A fitting algorithm for Markov-modulated Poisson processes having two arrival rates. *European Journal of Operational Research*, 29:370–377, 1987.

STD-1390D MIL. Level Of Repair Analysis (LORA). Technical report, Military Standard, United States Department of Defense: MIL-STD-1390D, 1993.

MIL-HDBK-965. Military handbook - acquisition practices for parts management (mil-hdbk-965), 1996.

MIL-PRF-49506. Logistics management information (mil-prf-49506), 1996.

MIL-STD-1390D. Level of repair analysis (mil-std-1390d), 1993.

MIL-STD-3018. Parts management (mil-std-3018), 2011.

S. Minner. Multiple supplier inventory models in supply chain management: A review. *International Journal of Production Economics*, 81-82:265–279, 2003.

K. Moinzadeh and C.P. Schmidt. An $(S-1, S)$ inventory system with emergency orders. *Operations Research*, 39(3):308–321, 1991.

C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

B. Moreta and I. Ziedins. Admission controls for Erlang's loss system with service times distributed as a finite sum of exponential random variables. *Journal of Applied Mathematics & Decision Sciences*, 2(2):119–132, 1998.

K. Morton. Bounds on the solution of the lagged optimal inventory equation with no demand backlogging and proportional costs. *SIAM Review*, 11(4):572–596, 1969.

K. Morton. The near-myopic nature of the lagged-proportional-cost inventory problem with lost sales. *Operations Research*, 19:7–11, 1971.

J.A. Muckstadt. A model for a multi-item, multi-echelon, multi-indenture inventory system. *Management Science*, 20(4):472–481, 1973.

J.A. Muckstadt. *Analysis and Algorithms for Service Part Supply Chains*. Springer: Berlin, 2005.

S. Nahmias. *Production and Operations Analysis*. McGraw-Hill, 6 edition, 2009.

B.L. Nelson and I. Gerhardt. On capturing dependence in point processes: Matching moments and other techniques. *Working Paper*, 2010. URL `http://users.iems.northwestern.edu/~nelsonb/Publications/GerhardtNelsonSurvey.pdf`.

W. Nelson. *Applied Life Data Analysis*. John Wiley & Sons, 1982.

W. Nelson. *Accelerated Testing*. John Wiley & Sons, 1990.

R. Oliva and R. Kallenberg. Managing the transition from products to services. *International Journal of Service Industry Management*, 14(2):160–172, 2003.

K.B. Öner, R. Franssen, and G.P. Kiesmüller. Life cycle costs measurement of complex systems manufactured by an engineer-to-order company. In R.G. Qui, D.W. Russel, W.G. Sullivan, and M. Ahmad, editors, *The 17th International Conference on Flexible Automation and Intelligent Manufacturing*, pages 569–589, 2007.

K.B. Öner, G.P. Kiesmüller, and G.J. Van Houtum. Optimization of component reliability in the design phase of capital goods. *European Journal of Operational Research*, 205:615–624, 2010.

P. Paterson, G.P. Kiesmüller, R. Teunter, and K. Glazebrook. Inventory models with lateral transshipments: A review. *European Journal of Operational Research*, 210 (2):125 – 136, 2011.

N.M. Paz and W. Leigh. Maintenance scheduling: Issues, results and research needs. *International Journal of Operations and Production Management*, 14(8): 47–69, 1994.

F. Pérès and J.C. Grenouilleau. Initial spare parts supply of an orbital system. *Aircraft Engineering and Aerospace Technology*, 74(3):252–262, 2002.

M.L. Pinedo. *Planning and Scheduling in Manufacturing and Services*. Springer, 2009.

M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 1994.

D.F. Pyke. Priority repair and dispatch policies for reparable-item logistics systems. *Naval Research Logisitics*, 37(1):1–30, 1990.

W. Romeijnders, R. Teunter, and W. van Jaarsveld. A two-step method for forecasting spare parts demand using information on component repairs. *European Journal of Operational Research*, 220:386–393, 2012.

S.M. Ross. *Stochastic Processes*. Wiley, 2 edition, 1996.

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

N. Safaei, D. Banjevic, and A.K.S. Jardine. Workforce constrained maintenance scheduling for military aircrapft fleet: a case study. *Annals of Operations Research*,

186:295–316, 2011.

R. Schassberger. *Warteschlangen*. Springer, 1973.

C.A. Schneeweiss. Hierarchical planning in organizations: Elements of a general theory. *International Journal of Production Economics*, 56-57:547–556, 1998.

C.A. Schneeweiss. Distributed decision making - a unified approach. *European Journal of Operational Research*, 150:237–252, 2003.

C.A. Schneeweiss and H. Schröder. Planning and scheduling the repair shops of the deutsche lufthansa ag: A hierarchichal approach. *Production and Operations Management*, 1(1):22–33, 1992.

C.A. Schneeweiss and K. Zimmer. Hierarchical coordination mechanisms within the supply chain. *European Journal of Operational Research*, 153:687–703, 2004.

G.D. Scudder. An evaluation of overtime policies for a repair shop. *Journal of Operations Management*, 6(1):87–98, 1985.

G.D. Scudder. Scheduling and labour assignment policies for a dual-constrained repair shop. *Intenational Journal of Production Research*, 24(3):623–634, 1986.

G.D. Scudder and R.C.H. Chua. Determining overtime policies for a repair shop. *Omega*, 15(3):197–206, 1987.

A. Sheopuri, G. Janakiraman, and S. Seshadri. New policies for the stochastic inventory control problem with two supply sources. *Operations Research*, 58(3): 734–745, 2010.

C.C. Sherbrooke. METRIC: A multi-echelon technique for recoverable item control. *Operations Research*, 16(1):122–141, 1968.

C.C. Sherbrooke. VARI-METRIC: Improved approximations for multi-indenture, multi-echelon availability models. *Operations Research*, 34(2):311–319, 1986.

C.C. Sherbrooke. *Optimal inventory modeling of systems: Multi-echelon techniques*. Wiley, 2 edition, 2004.

E.A. Silver, D.F. Pyke, and R. Peterson. *Inventory management and production planning and scheduling*. John Wiley & Sons, 1998.

F.M. Slay and C. Sherbrooke. The nature of the aircraft component failure process. Technical Report IR701R1, Logistics Management Institute, Washington D.C., 1988.

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

A. Sleptchenko, M.C. van der Heijden, and A. van Harten. Using repair priorities to reduce stock investment in spare part networks. *European Journal of Operational Research*, 163:733–750, 2005.

A. Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 1776.

J.S. Song and P. Zipkin. Inventory control in a fluctuating demand environment. *Operations Research*, 41(2):351–370, 1993.

J.S. Song and P. Zipkin. Inventories with multiple supply sources and networks of queues with overflow bypasses. *Management Science*, 55(3):362–372, 2009.

J.M. Spitter. *Rolling schedule approaches for supply chain operations planning*. PhD thesis, Eindhoven University of Technology, 2005. `http://alexandria.tue.nl/extra2/200511140.pdf`.

J.M. Spitter, C.A.J. Hurkens, A.G. De Kok, J.K. Lenstra, and E.G. Negenman. Linear programming models with planned lead times for supply chain operations planning. *European Journal of Operational Research*, 163:706–720, 2005.

D.H. Stamatis. *Failure mode and effect analysis*. ASQC Quality press, 1995.

D. Stoneham. *Maintenance Management and Technology Handbook*. Elsevier advanced technology, 1998.

R. Teunter and L. Duncan. Forecasting intermittent demand: a comparative study. *Journal of the Operational Research Society*, 60:321–329, 2009.

R.H. Teunter and L. Fortuin. End-of-life service. *International Journal of Production Economics*, 59:487–497, 1999.

R.H. Teunter and W.K. Klein Haneveld. The 'final order' problem. *European Journal of Operational Research*, 107:35–44, 1998.

H.G.H. Tiemessen and G.J. Van Houtum. Reducing costs of repairable inventory supply systens via dynamic scheduling. *International Journal of Production Economics*, 143:478–488, 2012.

G. Van Dijkhuizen and A. Van Harten. Optimal clustering of frequency-constrained maintenance jobs with shared set-ups. *European Journal of Operational Research*, 99:552–564, 1997.

K. van Donselaar, T. de Kok, and W. Rutten. Two replenishment strategies for the lost sales inventory model: A comparison. *International Journal of Production Economics*, 46-47:285–295, 1996.

J.P.J. Van Kooten and T. Tan. The final order problem for repairable spare parts under condemnation. *Journal of the Operational Research Society*, 60:1449–1461,

2009.

A.J. Van Weele. *Purchasing and supply chain management*. London: Cengage, 5th edition, 2010.

S. Veeraraghavan and A. Scheller-Wolf. Now or later: a simple policy for effective dual sourcing in capacitated systems. *Operations Research*, 56(4):850–864, 2008.

K. Vernooij. An aggregate planning for preventive maintenance of bogies by NedTrain. Master's thesis, Eindhoven University of Technology, School of Industrial Engineering, the Netherlands, 2011. `http://alexandria.tue.nl/extra2/afstversl/tm/Vernooij_2011.pdf`.

J. Verrijdt, I. Adan, and T. de Kok. A trade off between emergency repair and inventory investment. *IIE Transactions*, 30:119–132, 1998.

I. Vliegen. *Integrated planning for service tools and spare parts for capital goods*. PhD thesis, Eindhoven University of Technology, School of Industrial Engineering, 2009.

H.M. Wagner, R.J. Giglio, and R.G. Glaser. Preventive maintenance scheduling by mathmatical programming. *Management Science*, 10(2):316–334, 1964.

W. Wang and A.A. Syntetos. Spare parts demand: Linking forecasting to equipment maintenance. *Transportation Research Part E: Logistics and Transportation Review*, 47(6):1194 – 1209, 2011.

A.S. Whittmore and S.C. Saunders. Optimal inventory under stochastic demand with two supply options. *SIAM journal of applied mathematics*, 32(2):293–305, 1977.

V.C.S. Wiers. The relationship between shop floor autonomy and aps implementation success: evidence from two cases. *Production Planning & Control*, 20(7):576–585, 2009.

T.R. Willemain, C.N. Smart, and H.F. Schwarz. A new approach to forecasting intermittent demand for service part inventories. *International Journal of Forecasting*, 20:375–387, 2004.

R. Wise and P. Baumgartner. Go downstream: The new profit imperative in manufacturing. *Harvard Business Review*, 77:133–141, 1999.

T. Yoshihara, S. Kasahara, and Y. Takahashi. Practical time-scale fitting of self-similar traffic with Markov-modulated Poisson process. *Telecommunication Systems*, 17(1-2):185–211, 2001.

E.J. Zarybnisky. *Maintenance Scheduling for Modular Systems - Models and Algorithms*. PhD thesis, Sloan School of Management, Massachusetts Institute of Technology, 2011.

P. Zipkin. Old and new methods for lost-sales inventory systems. *Operations Research*, 56(5):1256–1263, 2008a.

P. Zipkin. On the structure of lost-sales inventory models. *Operations Research*, 56 (4):937–944, 2008b.

P.H. Zipkin. *Foundations of inventory management*. McGraw-Hill, 2000.

Summary

Spare Parts Planning and Control for Maintenance Operations

Interchangeable parts have revolutionized modern manufacturing. However, the idea of interchangeable parts was originally a maintenance innovation. Equipment that represents a significant financial investment (e.g. aircraft, rolling stock and MRI scanners) is usually maintained by replacing parts in need of maintenance with ready-for-use parts. In this manner, downtime of equipment due to maintenance can be kept to a minimum. To make this system work, it is crucial to have the right amount of spare parts available. This thesis studies the planning and control of spare part supply chains that support maintenance operations of the type described above. We identify three main types of spare parts:

- Rotables - These are items that constitute a sufficiently large subsystem of the original equipment to warrant a separate usage based maintenance strategy. Rotables are individually tracked and traced so that the correct usage can be ascribed to each rotable individually. Usually, there are dedicated resources for the maintenance and overhaul of rotables. Examples include aircraft engines, rolling stock bogies, and elaborate weapon or radar systems on frigates.
- Repairables - These are items that are repaired after replacement after which they are *ready-for-use* (RFU) again. Contrary to rotables, repairables do not have their own usage based maintenance strategy, and they are usually not individually tracked and traced. A repair shop handles the repair of many different types of repairables. Examples of repairables include compressors and pumps.
- Consumables - These are items that are discarded after replacement and bought new from a supplier. Generally these are relatively cheap items such as gaskets,

filters, and breakpads.

Demand intensity for rotable and repairable parts fluctuates over time because maintenance fluctuates over time due to maintenance programs and equipment degradation.

General framework

Chapter 2 presents a general framework for the planning and control of spare part supply chains. This framework outlines the decision functions at strategic, tactical, and operational level needed to effectively control a spare parts supply chain. It also describes the interactions and (hierarchical) relations between these decisions and provides an outline of how these decisions can be decomposed. As such, the framework is a type of taxonomy of different decision functions and their interrelations. Chapter 2 also presents a review of literature on decision support models for the various identified decision functions.

Rotables overhaul and supply chain planning

Chapter 3 studies the scheduled usage based maintenance of rotable parts. Rotable maintenance is subject to a usage based maintenance strategy which means that a rotable should not be used any longer than the maximum inter-overhaul time (MIOT). Traditional approaches to scheduling usage based maintenance focus on postponing overhaul as long as possible to take advantage of the technical life of the rotable. Chapter 3 takes a more direct approach to this problem by considering the costs of material and overhaul capacity over the entire lifetime of the equipment. We also integrate the scheduling of different rotable types because they share the same capacity for overhaul. The problem of scheduling rotable overhaul subject to capacity and material availability constraints for all rotable types that share the same capacity is formulated as a mixed integer linear program This problem is strongly $\mathcal{NP}$-hard, but computational evidence suggests that the provided MIP formulation can be used to solve real life instances. Furthermore, the linear programming relaxation is quite tight and can be used for sensitivity analyses. The main managerial insight is that it is inappropriate to focus on minimizing early overhaul. Instead, managers should focus on smoothing the overhaul workload of types of repairable jointly. This approach leads to significant savings in capacity costs.

Repairable expediting and stocking

Repairable spare parts are expensive and in many practical situations, it is not possible to buy new repairables at will. It is economically attractive to purchase these repairables jointly with the equipment because repairables are produced in larger series when they are also used to build new equipment. Furthermore, these repairables might not be available in the market at a later time. Demand for repairable items typically fluctuates over time, reflecting the fluctuating need for maintenance over time. Companies anticipate these demand fluctuations by leveraging the possibility of expediting the repair of defective parts, rather than buying new parts. Expediting repair incurs additional costs (compared to regular repair) either because an external repair shop charges extra or because an internal repair shop needs to adapt its operations to accommodate expedited repairs.

Chapter 4 studies the situation described above and supports two decisions at the tactical and operational level respectively: (i) How many repairable spare parts should the firm buy? and (ii) When should the firm request that the repair of a part is expedited? The first decision is at a tactical level and the second is at an operational level. Both are modeled by a single item stochastic inventory model. The fluctuations of demand over time are modeled by a Markov modulated Poisson demand process, and the possibility to expedite is modeled through two modes of inventory replenishment with different lead times. The shorter of these lead times is called the expedited lead time. For a fixed number of spare repairables and lot-for-lot replenishment, the optimal expediting policy may take two forms. The first form is simply to never expedite repair. The second form is a state dependent threshold policy, where the threshold depends on both the state of the modulating chain of demand and the pipeline of repair orders. Which type of policy is optimal can be determined by evaluating a simple closed form expression involving cost and lead time parameters.

Unfortunately the determination of optimal expediting policies as well as the decision how many repairables to buy suffer from tractability issues in general. Therefore Chapter 4 also presents a simple heuristic to determine a good expediting policy in combination with an amount of repairables to buy. This heuristic is based on results about optimal policies. In a numerical study involving a large test bed, this heuristic has an average and maximum optimality gap of 0.15% and 0.76% respectively.

Finally, Chapter 4 investigates the value of anticipating demand fluctuations by comparing optimal joint stocking and expediting optimization against naive heuristics that do not explicitly model demand fluctuations, or that separate the stocking and expediting policy decisions. These naive heuristics have optimality gaps of 12% on average and range up to 64% in our numerical work. The comparison with these naive

heuristics show that: (i) There is great value in leveraging knowledge about demand fluctuations, in making repair expediting decisions; (ii) Fluctuations of demand and the possibility to anticipate these through expediting repairs should be considered explicitly when deciding how many repairables to buy and can lead to substantial savings.

Chapter 5 extends the model of Chapter 4 to a multi-item multi-fleet multi-repair capacity setting. The model seeks to minimize the investment in all different types of repairable spare parts subject to constraints on the mean number of backorders for each fleet and constraints on the repair expediting load experienced by each repair capacity.

This last constraint allows the model to capture essential characteristics of smart scheduling policies, namely that the repair lead time can be shortened for parts that are in short supply and lengthened for parts that are in ample supply. The merit of the model in chapter 5 is that it can do this in a tractable manner. Even so, this optimization model is a non-linear non-convex integer programming problem. Lower bounds for this problem can be found through a column generation algorithm in which the pricing problem is exactly the problem studied in Chapter 4. A good feasible solution can be found within reasonable time using binary programming techniques. In extensive numerical experiments, the feasible solution we found had an optimality gap of 0.67% on average and at most 6.76%.

The same numerical study also quantifies the effect of considering repair shop flexibility through expediting compared to models in which stocking decisions are based on a single lead time. Explicitly considering these flexible lead times through expediting leads to an average reduction in repairable spare parts investment of 25% compared to the approach based on a single lead time for a large test-bed.

Consumables

Chapter 6 studies base-stock policies for consumables that are reviewed periodically. When the stock for consumables is depleted, it is a common procedure to use an emergency supply source to replenish the part almost instantaneously so that maintenance is not halted for lack of a part. All items that are replenished by the emergency procedure are lost to the normal mode of replenishment. This problem is mathematically equivalent to the classical lost sales inventory problem which consists of a periodically reviewed stock point that faces stochastic demand and loses any demand in excess of on-hand stock. Replenishment orders arrive after a deterministic lead time τ. At the end of each period, costs for lost sales and holding inventory are charged. For such systems, we are interested in minimizing the long run average cost per period.

The structure of the optimal policy for lost sales inventory systems with a positive replenishment lead time is still not completely understood, and the computation of optimal policies suffers from the curse of dimensionality as the state space is τ-dimensional. Base-stock policies are asymptotically optimal as the lost sales penalty costs approach infinity. However, computing the best base-stock policy for a lost-sales inventory problem efficiently remains a challenge. Chapter 6 presents an approximate method to compute the average cost rate under a given base-stock level. This approximation is based on several limiting results that have good convergence properties. Furthermore, this approximation satisfies Little's law with respect to the queue of pipeline orders. A simple heuristic to find the best base-stock level is to minimize the costs based on this approximation. A numerical study demonstrates that this heuristic has a cost performance of within 0.1% from the best base-stock policy on average and never more than 1.3% from the best base-stock policy across a large test bed.

About the author

Joachim Arts was born in Eindhoven on May 14, 1983. He completed grammar school at the Christiaan Huygens College in Eindhoven in 2001. After that, he worked as a train steward for one year in order to gather funds to serve as a missionary of the church of Jesus Christ of latter-day saints. He was called to serve in the West-Indies mission from 2002-2004 and spent time in Barbados, St. Martin and Surinam.

In 2007, Joachim obtained a BSc in Industrial Engineering and Management Science and in 2009 he received a MSc in Operations Management and Logistics. Both degrees were obtained cum laude from the Eindhoven University of Technology. His BSc thesis studies allocation rules for assemble-to-order systems and was supervised by dr. Kai Huang. His Master thesis studies dual sourcing inventory systems and was supervised by prof.dr. Gudrun Kiesmüller. This thesis was awarded with the price for the best master thesis in the field of operational research defended in the Netherlands in 2009 by the Dutch society for statistics and operational research.

From September 2009 to August 2013 he was a PhD student at the Eindhoven University of Technology under the supervision of prof.dr.ir. Geert-Jan van Houtum. During his PhD, Joachim visited MIT for 4 months to work with prof.dr. Retsef Levi.

As of september 2013, Joachim is working as an assistant professor at the Eindhoven University of Technology.