Proactive lateral transshipments and stock allocation via transient behavior of loss systems

Pielage, K.M.P.

Award date:
2018
Proactive lateral transshipments and stock allocation via transient behavior of loss systems

by
Karin Pielage

BSc Applied Mathematics
Eindhoven University of Technology 2015

Student identity number 0811447

in partial fulfilment of the requirements for the degree of

Master of Science
in Industrial and Applied Mathematics
and Operations Management and Logistics

Supervisors:
prof.dr.ir. O.J. Boxma, TU/e, SPOR
prof.dr.ir. G.J.J.A.N. van Houtum, TU/e, OPAC
prof. dr. F.C.R Spieksma, TU/e, DMA
ir. M.A Kuipers, ASML
ir. R.J.M van Sommeren, ASML
Subject headings: Spare parts management, Operational level planning, Finite contract duration, Proactive lateral transshipment, Stock allocation, Erlang loss model, Transient behavior, Heuristics
ABSTRACT

Motivated by service level agreements (SLA) in spare parts networks, proactive lateral transshipment and smarter stock allocation decisions are investigated as a way to comply with finite-horizon contracts. We consider a multi-item, multi-location, two-echelon spare parts model, consisting of one global warehouse and multiple local warehouses. Each local warehouse faces Poisson demand and is replenished by the global warehouse, where we assume exponentially distributed replenishment leadtimes. When new stock arrives in the global warehouse, this stock is allocated into the field. On operational level, decisions for these allocations as well as for proactive redistribution of parts in the field should be made with the goal of decreasing the total expected costs. This is done by reducing the risk on stockouts in order to decrease the total expected costs. Two heuristics are developed to serve this objective. In practice, allocation-only policies are often used for operational level decisions. Hence, one heuristic only makes allocation decisions, whereas the other also allows for proactive lateral transshipments.

By modeling each local warehouse as an Erlang loss system, the transient behavior of each local warehouse individually is studied. This behavior is used to find the distribution for the number of stockouts, which enabled us to formulate two Mixed Integer Non-Linear Programs (MINLP) with the objective to find the (trans)shipments that minimize the total costs until the end of the contract period. For small instances, the performance of the proposed heuristics is compared with two allocation-only heuristics: the FCFS heuristic and the heuristic that is currently used by ASML. The ASML heuristic and the two proposed heuristics perform significantly better than FCFS, with average cost reductions of 4.8%, 5.7% and 12.5%. More than 85% of this cost reduction is due to reducing the contract violation penalty costs. The total cost reduction is the largest for low failure rates, where the most advanced heuristic leads to savings of 42.5%. Although these results hold only for small instances, it shows great potential of using smart allocation rules and proactive lateral transshipments for finite-horizon contracts.
Before you lies the result of nine months of hard work. This thesis marks not only the end of my Master Thesis project conducted at ASML and the Eindhoven University of Technology as partial fulfillment of the masters Industrial and Applied Mathematics, and Operations Management and Logistics, but it also marks the end of my six-year student life. Completing this thesis would not have been possible without the help and support of a number of people, to whom I would like to express my gratitude.

First, I would like to thank my supervisors at the university, Onno Boxma and Geert-Jan van Houtum. Your suggestions and constructive feedback have been absolutely invaluable. Thank you Onno, for your ideas and discussions concerning the mathematical aspects of the thesis, and for providing me with the right tools to successfully complete this thesis. I wish you all the best on your retirement. Thank you Geert-Jan, for your guidance through this process, for steering me in the right direction, and for your critical view and for the insightful and constructive meetings that greatly assisted the research. Additionally, I thank Frits Spieksma for his willingness to complete my thesis assessment committee.

From ASML, I would like to thank Ruud van Sommeren for the opportunity to carry out this work in your department. Special thanks go out to Michiel Kuipers, my daily supervisor, for your support and advise, for keeping realistic goals in mind, and for giving me lots of freedom in conducting this research. I would also like to thank the team for their interest and for making my time at ASML a great one.

Next, I would like to thanks all my friends I met during my study for the study sessions at table 14, for the many coffee (and tea) breaks, and for all the laughter, silliness and fun we had in the last six years. Furthermore, a special thanks goes out to everyone who has helped me in such a way that they deserved a spot in my preface.

Finally, I would like to express my very great appreciation to my family for their continuous support and encouragement throughout my years of study. This accomplishment would not have been possible without them. Last but not least, I want to thank my boyfriend, Thomas, for his love and support. Thank you for listening to my struggles, for your advice, your patience and your faith. Thank you for everything.

Karin Pielage
Executive Summary

Capital goods are complex technical systems that are important for businesses as they are used to produce functional goods for their customers or for the delivery of valuable services. However, this kind of machinery is often highly capital intensive and simultaneously the bottleneck in the production process. To avoid high downtime costs, businesses owning capital therefore require high availability of their systems. Maximizing the availability by quickly identifying and replacing the failed components is key. It is common practice to put service level agreements (SLA) in place that assure quick responses such that these high availability requirements are attained at the end of the contract. However, the contract performance requirements should also be met over intermediate contract periods. Penalties may be incurred if these service requirements are not attained, which makes it crucial to track the service levels real-time during the contract period. To avoid these penalties, spare parts could be placed in stock close to the customer sites. However, these spare parts often are expensive and good management is required. On tactical level, the number of parts on stock are determined to avoid stockouts. The challenge is to minimize costs, which include costs for holding inventory and for replenishing, while still providing the promised service level. However, despite excellent tactical planning, stockouts still occur and need to be resolved at short notice. Therefore, in addition to tactical planning level actions, companies need to take many day-to-day operational planning level actions that assist good spare part management. These actions include reactive lateral transshipments in case of stockouts, proactive lateral transshipments and stock allocation decisions. Decisions for reactive lateral transshipments are often fixed as they depend on the distances between warehouses. On the contrary, decision rules for proactive lateral transshipments and stock allocation can be more dynamic. However, such dynamic rules are rarely seen in practice. The question arises on how to respond to the current situation with the resources and information available. Therefore, this thesis assists in answering two major questions:

1. When and how should stock be proactively transshipped between the local warehouses in anticipation of contract violations? (redistribution)

2. How should stock that becomes available in the global warehouse, repaired or new-buy, be allocated among the local warehouses? (allocation)

In this research, a multi-item, multi-location, two-echelon spare parts model is considered. The model consists of one global warehouse and multiple local warehouses around the world. Both warehouse types use a basestock policy for ordering parts at the supplier and the global warehouse, respectively. This system is a mixture of a backorder and an adjusted lost sales model, including
reactive lateral transshipments and emergency shipments. The local warehouses are grouped by regions and local warehouses within the same region can provide spare parts to each other in case of stockouts, i.e. the reactive lateral transshipments, for which static rules are put in place. For each local warehouse a SLA is set up that assures a certain service level within a fixed and finite period of time, called the contract period. The SLAs for each local warehouse are defined by the number of stockouts that are allowed during the contract period. Spare parts can be send proactively between local warehouses in order to reduce the risk on violating the SLA. Furthermore, defective parts trigger an order at the supplier, and when these parts arrive at the global warehouse, stock is allocated to the local warehouses that have the highest need.

The possibilities to reduce the overall costs by increasing the contract performances using smarter decisions are investigated. More specific, it is investigated how smarter allocation rules and proactive lateral transshipments can be used to reduce the expected costs until the end of the contract period. This is done by reducing the risk of possible stockouts in order to increase the probability of meeting an SLA. The goal is to present a decision model for the decisions for proactive lateral transshipments and stock allocation based on the current state of the network.

For the development of two heuristics that serve this objective, we resort to transient loss systems, as this spare parts network can easily be translated to a queueing system, where finite time periods can be considered. These transient loss systems are used to find the distribution for the number of stockouts, which is used to calculate the probability of violating contracts. Based on these probabilities, a cost function for the expected costs until the end of the contract period is derived. We formulate two optimization problems that minimize this cost function by making proactive lateral transshipments and allocation decisions.

In practice, allocation-only policies are often used in the operational level decisions. Therefore, the first of our two heuristics, called PASA, only allows for allocation decisions. Our second heuristic, PASA-P, is an extension of the first by also allowing proactive lateral transshipments. These two heuristics are performance anticipating heuristics that take the current state of the system into account, e.g. the on-hand inventory levels in comparison to the basestock levels, the remaining number of allowed stockouts and the time until the end of the contract period.

The performance of our proposed heuristics is compared with the FCFS heuristic and the NORA heuristic, which is currently used by ASML, where decisions are based on the steady-state Erlang blocking probability. The comparison of the performance is done for small networks and with a limited number of instances due to the computational complexity of the two proposed heuristics. We found that the NORA, PASA and PASA-P heuristics showed a significant improvement compared to the FCFS heuristic.

In this small network, the average cost per contract period for the FCFS heuristic is $2.39 \cdot 10^5$ euro, of which 73% is due to penalty costs. The NORA, PASA and PASA-P heuristics are able to reduce this average costs by 4.8%, 5.7% and 12.5%, respectively. The main contributor to these reductions
is the reduction of penalty cost, which is good for more than 85% of the cost reduction. The total cost reduction is the largest for low failure rates, where PASA-P leads to a cost reduction of 42.5%. Although these results hold only for small instances, it shows great potential for using smarter allocation methods and proactive lateral transshipments to achieve cost reductions in large-scale systems.

We can conclude that, although NORA already showed a significant improvement over FCFS, the potential of reducing cost in large scale systems of these two proposed heuristics should be acknowledged. Especially, the addition of proactive lateral transshipments in PASA-P seem to be of great value in reducing the average total costs, while simultaneously reducing the violated contracts.
# CONTENTS

Abstract i  
Preface ii  
Executive summary iii  

1 Introduction 1  

2 Literature review 6  
2.1 Finite horizon contracts 6  
2.2 Proactive lateral transshipments 7  
2.3 Stock allocation 8  
2.4 Discussion 9  
2.5 Transient loss systems 10  

3 Model description 12  
3.1 Allocation and redistribution problem 12  
3.2 Service level agreements 15  
3.3 Overview of assumptions 16  

4 Heuristic approach 18  
4.1 Isolated warehouses 18  
4.2 Distribution of the number of losses 19  
4.3 Probability of exceeding allowed stockouts 27  
4.4 Optimization problem 28  

5 Simulation design 32  
5.1 Heuristics 32  
5.2 Allowed number of stockouts 34  
5.3 Steps and procedures 36  

vi
Contents

6 Computational study 39
  6.1 Test beds and objectives .................................................. 39
  6.2 Numerical results .............................................................. 42

7 Conclusions and discussion 50
  7.1 Main results .................................................................. 50
  7.2 Limitations and recommendations .......................................... 51

Bibliography 53

Appendices 56

A Exact loss probabilities for a $M/M/1/1$ loss system 57

B Complexity of the probability generating function of a $M/M/2/2$ loss system 59

C Confidence intervals for the computational study 61
# List of Figures

1.1 An example of shipments in a multi-location spare parts network . . . . . . . . . . . . 2
1.2 Multi-location spare parts network under consideration . . . . . . . . . . . . . . . . 4

3.1 Transshipment time and cost flow . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

5.1 Simulation steps for each time step . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

6.1 Effect of varying the parameters in the test bed . . . . . . . . . . . . . . . . . . . . 44
6.2 Contribution of different cost factors as fraction of the total cost per contract period for simulation study 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 49
LIST OF TABLES

5.1 Overview of the four heuristics .................................. 33
6.1 Base parameter values for computational study .................. 40
6.2 Parameter settings the computational study ....................... 41
6.3 Summary of the computational results for simulation study 1 ........ 43
6.4 Contribution to total cost per contract period for simulation study 1 (in %) ........ 43
6.5 Summary of the number of violated contracts for simulation study 1 ........ 46
6.6 Summary of the computational results for simulation study 2 ........ 47
6.7 Summary of the number of violated contracts for simulation study 2 ........ 47
6.8 Average absolute costs due to different cost factors per contract period for simulation study 2 (\(10^4 \text{ €}\)) ....................... 48
C.1 The 95% confidence interval for each instance of simulation study 1 ........ 61
C.2 The 95% confidence interval for each instance of simulation study 2 ........ 64
Capital goods are complex technical systems that are important in the production of goods and services. Examples are aircrafts, advanced medical tools, baggage handling systems and lithography machines. These capital goods are important for businesses as they are used to produce functional goods for their customers and to deliver valuable services. However, this kind of machinery is often highly capital intensive and simultaneously the bottleneck in the production process. An undesirable consequence is that this combination could result in high downtime costs – which consists of lost production costs and liability costs. In order to avoid these downtime costs and loss of goodwill, businesses owning capital require high availability of their systems and failures should be resolved quickly.

Minimizing the delay by quickly identifying and replacing the failed components is key. This requires specialized knowledge of these high tech systems, which is only available at the original equipment manufacturer (OEM). Therefore, the OEMs do not only sell their systems, but often they also provide after-sales service conform a service level agreement (SLA). This SLA assures a certain service level within the given contract period. These contracts often span a few years, but there are also intermediate reviewing points at which the contract performance should be met. Throughout this thesis, the term ‘contract period’ will correspond to these shorter reviewing periods. Penalties may be incurred if these service requirements are not attained, which makes it crucial to track the service levels real-time during the contract period.

To avoid these penalties, spare parts could be placed in stock close to the customer sites, enabling quick responses to failures of parts. However, these spare parts, such as those used in aerospace maintenance, manufacturing facilities and power plants, are expensive and good management is required. Since an increasingly large amount of money is involved in spare parts management, most industries have spent a lot of effort in making spare parts management more (cost-)efficient over the past years. The challenge is to minimize costs, including costs for holding inventory and replenishing, while still providing the promised service level.

One method that has proven to be effective in cost reduction is the pooling of inventory, e.g. by Tagaras (1999). Furthermore, Kranenburg and Van Houtum (2009) incorporate reactive lateral
transshipments in case of a stockout in their model. They show that the costs of spare parts provisioning can be up to 50% lower when using reactive lateral transshipments compared to not using these transshipments in their tactical planning. However, in practice, the customer locations and their associated local warehouses are often geographically dispersed all over the world. Additionally, most spare parts are classified as slow moving, i.e. low failure rate, and expensive items. Where the shipment time is relatively constant, (external) supply lead times are usually long and highly uncertain. Therefore, despite excellent tactical planning, stockouts still occur and need to be resolved at short notice.

In addition to tactical planning level actions, companies need to take many day-to-day operational planning level actions. These actions include reactive lateral transshipments, proactive lateral transshipments and stock allocation decisions: decisions for reactive lateral transshipments identify which local warehouse should support another local warehouse that experiences stockout; proactive lateral transshipments decisions specify when and how stock should be redistributed among the local warehouses to reduce the risk on stockouts; and stock allocation decisions determine which local warehouse should be replenished by the incoming stock in the global warehouse. These decisions are depicted in Figure 1.1.

At an operational level, also real-time information can be added to the decision making process. For reactive lateral transshipment, many OEMs have implemented a static allocation rule that fulfills a spare parts demand from the closest warehouse with a positive inventory level. Decision rules that are more dynamic have also been investigated extensively, but since most SLAs are time-based, the possible warehouses eligible for support are limited. As opposed to reactive lateral transshipments, decision rules for proactive lateral transshipments and stock allocation can be more dynamic. However, such dynamic rules are rarely seen in practice. Moreover, these decisions require the entire network of warehouses to be taken into account.
Allocation and proactive transshipment decisions are two decisions that must be taken in highly dynamic environments. These decisions are driven by the information available about the current state of the system, the current service levels and the information available about the demands that are likely to occur in the time periods that these decisions will affect. Operational planning in a distribution system involves, among others, determining the best way to respond to the current situation with the resources and information available.

Based on the situation described above, decision makers face two major questions:

1. When and how should stock be proactively transshipped between the local warehouses in anticipation of contract violations? (redistribution)

2. How should stock that becomes available in the global warehouse, repaired or new-buy, be allocated among the local warehouses? (allocation)

The objective of this thesis is to present an optimization model that assists decision makers with answering these two questions. More specifically, we intend to come up with heuristics that have the potential to perform well in practice and are focused on both proactive lateral transshipments and smart stock allocation. This should be a dynamic decision model that incorporates information about the real-time state of the network, such as the current stock levels, the contract performances and the remaining time left in the contracts.

This thesis was done in close collaboration with ASML. ASML is the world leader in the manufacturing of chip-making equipment. They design, develop, integrate, market, and service advanced lithography machines. The company offers an extensive portfolio of lithography machines that are used for the manufacturing of integrated circuits (also called chips). Nowadays, ASML operates in 16 countries and employs over 19,000 employees, of which more than 7,000 work in the research and development department (ASML, 2018a). Last year alone, ASML realized a turnover of more than 9 billion euros.

With machines that cost up to 100 million euros, it is obvious that downtime is very costly and that the performance of these contracts should be managed carefully (ASML, 2018b). ASML’s spare parts network consists of multiple local warehouses around the world, where spare parts are strategically positioned for when systems are down. This makes the network of ASML very suitable for investigating whether the allocation and redistribution decisions can improve the performance. As their network, however, is very large, the overarching objective of this thesis is to present heuristics that have the potential to reduce the overall costs of large-scale systems significantly by increasing performance.

Although this thesis will mainly consist of theoretical and general research, we wish to compare performance of this heuristic with heuristics used in practice. A logical choice for these heuristics is the heuristic that is currently used within ASML. It should be noted that this heuristic only makes
allocation decisions, but this provides an excellent opportunity to test the added value of proactive lateral transshipments.

The network that is studied in this thesis is a simplified version of the spare parts network of ASML: a multi-item, multi-location, two-echelon spare parts model. The model consists of one global warehouse and multiple local warehouses around the world. Both warehouse types use a basestock policy for ordering parts at the supplier and the global warehouse, respectively. This system is a mixture of an adjusted lost sales and a backorder model: demand at the local warehouses that cannot be satisfied directly is registered as lost, but can be satisfied via a reactive lateral transshipment or emergency shipment; replenishment orders from the local warehouses for parts that are not on-hand at the global warehouse are backordered. The local warehouses are grouped by regions and local warehouses within the same region can provide spare parts to each other in case of stockouts, the reactive lateral transshipments mentioned above. Furthermore, service parts can be sent proactively between local warehouses within the same region but also across different regions. Defective parts trigger an order at the supplier, which can be either for repair of the defective part or for the production of a new part. Upon a repair or production arrival in the global warehouse, the local warehouses are replenished by allocating the stock to the local warehouses that have the highest need. Figure 1.2 shows the conceptual model of the network under consideration. Here, static rules are put in place for reactive lateral transshipments, but decisions for stock allocations into the field and proactive lateral transshipments that minimize the total expected costs over all local warehouses at the end of the contract period should still be made, which is contained in the objective of this thesis.

![Multi-location spare parts network](image)

Figure 1.2: Multi-location spare parts network under consideration

With regard to scoping, there are several topics and concepts that are considered out of scope. First, tools are left out of scope and only spare parts are included. Tools planning is substantially different from spare parts planning, and therefore only the operational planning for parts is considered.
Second, upgrades of old machines and installs of new machines are out of scope. These activities require parts that are not included in the operational planning for the machines for which there is an SLA in place for the after-sales service. In terms of warehouses, we will only look at the global warehouse and the local warehouses downstream. Thus, suppliers itself will not be taken into account, but the supplier unreliability is still taken into account by modeling the supplier lead time as a random variable.

The main contributions of this thesis are summarized as follows:

1. We are the first to conduct research in the area of stock allocation decision with performance reviewing periods of finite length, of which the intersection in literature is empty. This empty intersection holds for the proactive lateral transshipment decisions as well and the combination of these (trans)shipments with finite horizon contracts is also included in this thesis.

2. Although both stock allocation and proactive lateral transshipment decisions have been studied extensively separately, only few papers have considered both decisions in one model. We contribute by not only including both decisions in one model, but also by dealing with service level agreements within fixed time windows in this model.

3. The heuristic implemented at ASML and our two proposed heuristics are compared with the static first-come first-serve heuristic for multiple small instances. The improvements for these three heuristics are significant. In a small network, the average cost per contract period for the FCFS heuristic is $2.39 \cdot 10^5$ euro, of which 73.5% is due to penalty costs. The NORA, PASA and PASA-P heuristics are able to reduce this average costs by 4.8%, 5.7% and 12.5%, respectively. The main contributor to these reductions is the reduction of penalty cost, which is good for more than 85% of the cost reduction. The total cost reduction is the largest for low failure rates, where PASA-P leads to a cost reduction of 42.5%. Although these results hold only for small instances, it shows great potential for using smarter allocation methods and proactive lateral transshipments.

The remainder of this thesis is organized as follows. Chapter 2 provides an overview of the related literature and positions the contribution of this work with respect to this literature. In Chapter 3, the formal description of the redistribution and allocation problem is presented. In Chapter 4, a heuristic approach is presented and optimization problems are formulated for the allocation-only problem as well as for the problem including proactive lateral transshipments. The design of the simulation is discussed in Chapter 5. Chapter 6 presents a computational study to evaluate the performance of these two heuristics in a test bed for small instances. Finally, the conclusion and recommendations for further research are given in Chapter 7.
Although spare parts inventory systems have been studied extensively, there is little attention for the incorporation of lateral transshipments in models with a finite horizon service contract. One method that has proven to be effective in reducing the overall costs of the spare parts inventory system, while still meeting all SLAs, is the use of reactive lateral transshipments. They are often used in practice, but they are not capable of preventing stockouts from occurring in the future. Furthermore, the often-used First-Come, First-Served (FCFS) allocation rule is not always optimal. Therefore, the objective of this literature review is to study the available literature on methods that use proactive lateral transshipments and smarter allocation rules to facilitate meeting finite horizon contracts for systems similar to the one of ASML.

Motivated by these contracts over a fixed review period, research has been done on the performance of such contracts in a finite horizon setting, as opposed to steady state performances. In this literature review we start with a brief discussion of these finite horizon contracts. However, proactive lateral transshipments and allocation policies are not taken into account in the literature and they must be investigated in more detail. Therefore, we focus on three other streams of literature. The first stream focuses on proactive lateral transshipment, whereas the second stream focuses on the allocation strategies. After investigating these two streams in detail, we come to the conclusion that this literature cannot assist in achieving our goal. Nevertheless, transient loss systems might offer us a way forward from the difficulties encountered in literature. An additional study is required and therefore, a third stream of literature is provided, diving more into these transient loss systems.

2.1 Finite horizon contracts

Chen et al. (2003) show that in a system under periodic review with zero lead time, the expected fill rate over a finite horizon is greater or equal to that over an infinite horizon for a given basestock level. The fill rate is defined as the percentage of demand that is immediately satisfied from stock at hand. This is due to potential double counting of backorders and order crossover. It is proven by Banerjee and Paul (2005) that the expected fill rate is non-increasing in the length of the contract
2.2 Proactive lateral transshipments

In most industries, a large amount of money is involved in the spare parts management, which makes efficient management crucial. More flexible systems allow lateral transshipments within an echelon. Optimal lateral transshipment policies are studied intensively and the reader is referred to Wong et al. (2006), Paterson et al. (2011) and Topan et al. (2018) for extensive reviews on this subject. The most relevant insights are incorporated in the remainder of this section.

With regards to redistribution of goods to prevent stockouts, the redistribution policy proposed by Bertrand and Bookbinder (1998) adjusts stock to achieve equal marginal cost over all retailers just before the replenishment period. It is one of the first papers that allow for multiple locations with nonidentical costs, and with non-zero order leadtime and only the costs for the remainder of the cycle are minimized. In a similar setting, Tagaras (1999) proposes three transshipment policies for a system with three locations that have identical costs. He shows that the model that attains the lowest total expected cost is the one taking the next period into consideration. Furthermore, it was found that it is beneficial to create balanced pooling groups that face similar demand.

Two other preventive lateral transshipment policies, based on availability (TBA) and inventory equalization (TIE), are proposed by Banerjee et al. (2003). TBA redistributes stock to prevent stockouts in the next period, whereas TIE reduces the number of stockouts over the long run. They use the basestock level, the current stock level and the demand rate for their decisions, but

period. Thomas (2005) focuses more on the distribution of the fill rate rather than the expectation. He investigates the effects of the horizon length, the demand distribution and the desired probability to meet the fill rate target, and concludes that the review length can both play in favor or against the manager and customer. Katok et al. (2008) also investigate the effects of the length of the review period on the stocking decisions. They use laboratory experiments and show that a longer horizon is more effective in inducing higher service performances than a shorter horizon. This is in contrast with the findings of Thomas (2005), which is most likely due to the influence of feedback in these experiments. More recently, Chen and Thomas (2018) tested several rules for allocation for a single item, after demand has been realized.

Unlike the aforementioned, Al Hanbali and Van der Heijden (2013) consider the interval availability as a service measure. They conclude that the expected availability in an interval of finite length is equal to the expected availability over an infinite horizon.

Proactive lateral transshipments are not taken into account in any of the aforementioned articles, while allocation was only considered by Chen and Thomas (2018), and none of them consider a multi-item model with multiple review periods in one contract period. Hence, in what follows, proactive lateral transshipments and allocation policies are investigated more extensively.

2.2 PROACTIVE LATERAL TRANSSHIPMENTS

In most industries, a large amount of money is involved in the spare parts management, which makes efficient management crucial. More flexible systems allow lateral transshipments within an echelon. Optimal lateral transshipment policies are studied intensively and the reader is referred to Wong et al. (2006), Paterson et al. (2011) and Topan et al. (2018) for extensive reviews on this subject. The most relevant insights are incorporated in the remainder of this section.

With regards to redistribution of goods to prevent stockouts, the redistribution policy proposed by Bertrand and Bookbinder (1998) adjusts stock to achieve equal marginal cost over all retailers just before the replenishment period. It is one of the first papers that allow for multiple locations with nonidentical costs, and with non-zero order leadtime and only the costs for the remainder of the cycle are minimized. In a similar setting, Tagaras (1999) proposes three transshipment policies for a system with three locations that have identical costs. He shows that the model that attains the lowest total expected cost is the one taking the next period into consideration. Furthermore, it was found that it is beneficial to create balanced pooling groups that face similar demand.

Two other preventive lateral transshipment policies, based on availability (TBA) and inventory equalization (TIE), are proposed by Banerjee et al. (2003). TBA redistributes stock to prevent stockouts in the next period, whereas TIE reduces the number of stockouts over the long run. They use the basestock level, the current stock level and the demand rate for their decisions, but
costs are not taking into account. Similar to the policies proposed by Bertrand and Bookbinder (1998) and Tagaras (1999), both policies have the disadvantage of not being able to respond to stockout before or after distribution.

A new policy is proposed by Lee et al. (2007), where transshipment decisions are based on the probability of no stockout during the remainder of the period. This policy was shown to perform better than the TIE and TBA policy, when the transportation costs are low, but not for increasing transportation costs. Meissner and Senicheva (2018) focus solely on the decisions for proactive lateral transshipments in a lost sales model. They use a dynamic programming approach to find the optimal proactive transshipment amount to maximize the profit of the whole network until the new order period. They resort to near-optimal solutions, among which is a modification of the TIE policy by Banerjee et al. (2003) that includes costs.

Tiacci and Saetta (2011) present a heuristic to decide when and how much to transship among retailers while minimizing the total expected costs until the next replenishment arrival from the central depot. This model considers only two retailers, such that the decision about receiving and sending retailers is eliminated. Feng et al. (2017) consider a similar model as Tiacci and Saetta (2011), but with multiple non-identical retailers, where the objective is to find the optimal timing and size of preventive transshipments. They use a decomposition method, a sorting heuristic and both forward and backward dynamic programming to find the optimal policy. In a continuous review system, proactive lateral transshipments additionally require a decision on the optimal timing for a transshipment. Seidscher and Minner (2013) present a Semi-Markov decision problem formulation for determining the optimal timing and source for proactive transshipments. Inventory changes are taken as decision epochs. Their exact analysis is limited to small networks.

Reactive and proactive transshipment together are considered by Zhao et al. (2008). Their production based model uses reactive lateral transshipments in case of stockouts, but can also separately allocate stock when new inventory is produced. Each location is modeled as a $M/M/1$ single-server, make-to-stock queueing system, but only two locations and one item are taken into account. Two other models that investigate the benefits of shipping additional stock in case of stockouts are the model of Paterson et al. (2012) and Glazebrook et al. (2015). They both exploit economies of scale and do this by using dynamic programming for finding the optimal policy. They both assume that this transshipment will be the last one ever made. Whereas Paterson et al. (2012) model a single-item system under continuous review in steady state, Glazebrook et al. (2015) model a multi-item backorder system under period review and in a finite horizon.

2.3 Stock allocation

Operational planning in a distribution system also involves determining the best way to allocate available inventories to the local warehouses. Static allocation rules, such as the First-Come First-
2.4 Discussion

Serve (FCFS) policy, are often not optimal and better strategies are investigated. De Véricourt et al. (2002) introduce a multilevel rationing policy and show that this outperforms FCFS in a make-to-stock system. They describe the model as a single-server, single-product, make-to-stock queue with multiple demand classes. It turns out that the optimal policy is a threshold policy.

Axsäter et al. (2002) propose two computationally tractable heuristics for a single-item backordering model: one is an ordering rule for warehouse replenishments and the other is a rule for allocating stock from the warehouse to the retailers. This model is based on a periodic review policy.

In contrast to the aforementioned papers, a multi-item system is considered by Caggiano et al. (2006). For a repair and distribution system with backorders, their model uses real-time information in deciding which items to repair, where to ship available units, and by what mode to ship them in each period of the planning horizon, based on a rolling horizon. They formulate their allocation problem as a large-scale linear program, which is solvable by part and therefore allows for making operational decisions in large-scale service parts distribution systems (Caggiano et al., 2006).

Foreman et al. (2010) extend the model of Caggiano et al. (2006) by capturing shipments, multiple expedited transportation modes and non-linear transportation cost, and by including transportation scheduling and capacity constraints. Furthermore, Caggiano et al. (2006) develop heuristic solutions, whereas Foreman et al. (2010) compute solutions to an approximate, linearized version of the model. Another paper based on the model of Caggiano et al. (2006) is the work of Topan and Van der Heijden (2018). In addition to Caggiano et al. (2006), they use both reactive and proactive transshipments and lateral transshipments are allowed. Furthermore, they focus on the operational decisions, by assuming fixed basestock levels, where the repair capacity is left out of scope. They consider a model with multiple types of transshipments and show that proactive emergency shipments contribute most to the downtime reduction.

Somarin et al. (2017) consider a repairable service parts inventory system with a central repair facility and several locations. They develop a cost effective after-repair service parts allocation policy, which minimizes operational costs and effectively fulfills demand. Their heuristic is based on the relative value function and average backorder cost at a single location. Results show that this heuristic policy outperforms the myopic policy.

2.4 Discussion

The first remark about the studied literature is that none of them include the current contract performance as a decision variable in their model. Topan et al. (2018) and Topan and Van der Heijden (2018) also identified this gap in the literature by stating that dealing with service level agreements within fixed time windows needs further investigation.
We found that most of the research on lateral transshipments is about reactive transshipments. There are only a few papers with both proactive and reactive transshipment, while there are even less that include a decision making policy for a combination of proactive transshipments and allocation of spare parts from the global warehouse. Furthermore, we conclude that the intersection between finite horizon review periods and proactive lateral transshipments, as well as the intersection between finite horizon review periods and stock allocation appears empty. To the best of our knowledge, no other paper has dealt with these intersection, and it therefore needs further investigation.

Regarding the modeling approach, many papers use a dynamic programming approach. Modeling the problem as a Markov decision process and then solving it using linear programming, policy iteration or value iteration is common in the lateral transshipment literature. However, for large networks with hundreds or thousands of spare parts, the state space would grow rapidly and the model would suffer from the curse of dimensionality.

To summarize, the goal is to come up with a method that will incorporate dynamic proactive transshipment and allocation decisions in a setting with finite horizon contract periods. Looking a finite time period ahead for systems that can be easily translated into a queueing system, it may look natural that transient behavior of queueing systems is a promising direction for further investigation. In the next section, research on these transient queueing systems is presented.

2.5 Transient loss systems

Motivated by telephone traffic, Takács (1962, Ch. 1) was among the first to investigate transient queueing models. He derived an expression for the time-dependent state probabilities of an $M/M/1/K$ system by using transforms. These state probabilities can also be used to derive the blocking probability for an $M/M/1/1$ system with no waiting room. Yunus (1990) presents an exact expression for the transient blocking probability in an $M/M/n/n$ loss system, where the initial probability distribution is arbitrary. Virtamo and Aalto (1998) develop a method for calculating the blocking probabilities in an $M/M/n/n$ system without the need to solve the probabilities of all the other states. They use the time-dependent state probabilities for an infinite system, but compensate with an additional term in the differential equations to translate the problem to a finite system. For a multi-server loss system, Abate and Whitt (1998) show how to compute the time-dependent blocking probabilities by numerically inverting the Laplace transform.

The papers above focus mainly on the transient blocking probability of queueing systems, but we are also interested in the number of losses within a given interval. In another chapter of Takács (1962), which focuses on teletraffic, not only the number of busy lines at time $t$ is investigated, but also an expression is found for the number of losses in an interval $(0, t)$, independent of the initial state. Ferrante (2009) presents an explicit form for the probability of no lost customers in $(0, t)$
for an $M/M/1/1$ system given the initial state of the system and an iterative procedure is given to determine the distribution of the total number of losses in $(0, t)$. While the work of Ferrante (2009) is inspired by the location problem of emergency vehicles (ambulances), we use it for an application in spare parts systems. This distribution for the number of losses was, however, only analyzed for an $M/M/1/1$ system, where we would like to find the distribution also for an $M/M/2/2$ system or even for general $M/M/K/K$ systems as this corresponds to a basestock level of 2 or even more.

In conclusion, the existing literature related to proactive lateral transshipments and allocation decisions, as a way to more easily comply with finite-horizon contracts, was investigated. Although these topics both have been studied extensively separately, only few papers have considered both decisions in one model. As emphasized in the introduction, we intend to come up with a heuristic approach that performs well in practice and focuses on both proactive transshipments and stock allocation. Here, the goal is to not exceed the number of allowed stockouts for each local warehouse. This is done by extending and generalizing the work of Ferrante (2018) for a $M/M/K/K$ (for $K \geq 2$) model to find the distribution for the number of stockouts, and use this as a decision variable in our model for proactive lateral transshipments and allocation decisions.
Chapter 3

Model Description

In this chapter, the model description is presented. First, a description of the two-echelon spare parts inventory system is provided and the notation that is used throughout this thesis is introduced. Subsequently, the service level agreements of this model are described. We conclude with summarizing the assumptions that are made for modeling this allocation and redistribution problem.

3.1 Allocation and Redistribution Problem

We consider a two-echelon spare parts inventory system consisting of a global warehouse and multiple local warehouses. Let \( J \) denote the (non-empty) set of warehouses, numbered \( j = 0, \ldots, |J| \). This set of warehouses consists of the global warehouse, denoted by \( J_g = \{0\} \), and the set of local warehouses, \( J_l = \{1, \ldots, |J| \} \) such that \( J_g \cup J_l = J \). Machines at each local warehouse consist of multiple critical components, which we refer to as Stock-Keeping Units (SKUs). Let \( I \) denote the (non-empty) set of SKUs; they are numbered as \( 1, \ldots, |I| \). Critical components are subject to failures. Upon a failure, the corresponding machine becomes non-operational. When the defective part is replaced by a new one, the machine is operational again. Each local warehouse is responsible for serving customers with one or more machines, whereas the global warehouse only serves for collecting all incoming supply and allocating these spare parts to the local warehouses. Machines are only located at local warehouses and hence, failures can only occur at these local warehouses. Failures for each SKU \( i \in I \) at local warehouse \( j \in J_l \) are assumed to occur according to a Poisson process with a constant rate \( \lambda_{i,j} \). This rate depends on the type and number of machines at the local warehouses. When a machine experiences a part failure, the part will request a new part from its designated local warehouse and if this local warehouse has stock on hand, it will immediately be provided by that local warehouse. The corresponding time and cost of the shipment from local warehouse to machine are both considered to be zero.

Two types of local warehouses are distinguished: main local warehouses and regular local warehouses, shortly referred to as mains and regulars, respectively. The subset of main local warehouses
is denoted by $K$, and all other local warehouses $j \in J_l \setminus K$ are regular local warehouses. The difference between these two types of local warehouses is that main local warehouses can be the supplier of a reactive lateral transshipment, whereas a regular local warehouse cannot. We assume that $|K| > 0$, so that these lateral transshipments can take place. Furthermore, we distinguish different regions in which the local warehouses are located. These regions often represent countries, but for large countries multiple regions can be defined. Let $R$ denote the set of all regions, numbered $r = 1, \ldots, |R|$. Let $r_j \in R$ denote the region in which local warehouse $j$ is located.

If local warehouse $j$ does not have stock on hand when a part request arrives, it tries to obtain the part from (another) main local warehouse in the same region, i.e. for some $k \in r_j, k \neq l$, via a reactive lateral transshipment. The corresponding reactive lateral transshipment time from main local warehouse $k \in J_l, k \neq j$ is $t_{rlt}^{jk}$ and the corresponding cost is $c_{rlt}^{jk}$. For these lateral transshipments, other local warehouses are checked in a predefined order. Reactive lateral transshipments are not performed across regions, so a local warehouse does not check all other local warehouses, but only checks $l_j$ other warehouses for $j \in J_l$. Define vector $\sigma(j) := (\sigma_1(j), \ldots, \sigma_{l_j}(j))$ as this sequence for local warehouse $j \in J_l$. Note that since reactive lateral transshipments can only take place within the same region, all local warehouses in $\sigma(j)$ must be located in region $r_j$.

Furthermore, let $K(k, \tilde{k})(\subset K)$ denote the subset of main local warehouses that are checked prior to main $\tilde{k}$ according to the pre-specified order $\sigma(k)$.

If neither the local warehouse itself nor one of the other local warehouses has stock on hand, an emergency shipment is requested from the global warehouse. The corresponding shipment time is $t_{em}^{j}(\leq t_{rlt}^{jk}, k \in J_l, k \neq j)$, and cost is $c_{em}^{j}(\geq c_{rlt}^{jk}, k \in J_l, k \neq j)$ for $j \in J_l$. We assume that emergency orders at the global warehouse can always be supplied, but can cause backorders at the supplier. Both $c_{rlt}^{jk}$ and $c_{em}^{j}$ are additional costs compared to the transshipment cost for an immediate part request fulfillment, i.e. from the assigned local warehouse to the machine group.

Part requests at the local warehouses trigger a request for new parts from the global warehouse. Consequently, the global warehouse orders this part at the supplier. This parts are replenished either by repairing the defective part or producing a new part and it takes exponentially distributed supply leadtime with mean $t_s^i$ for SKU $i$ to arrive at the global warehouse. We assume that the supplier has ample capacity for repairing defective parts and producing new parts. Upon arrival at the global warehouse, the global warehouse then has to make a decision to which local warehouses the parts will be sent, and this shipment time, called the allocation time, is denoted by $t_a^j$. The corresponding shipment cost is $c_a^j$. The global warehouse does not keep stock on-hand, so all incoming parts are allocated shortly after arriving.

Regardless of whether there has been a part request or not, parts can be transshipped between local warehouses, the so-called proactive lateral transshipments, to mitigate risk on stockouts and contract violation. Whereas reactive lateral transshipments cannot be performed between all warehouses, proactive lateral transshipments can. There is more time available for such a proactive
transshipment since they are not direct responses to failures, whereas reactive lateral transshipments do need a very quick response. The decision to make these transshipments can be made at any time, and the proactive lateral transshipment time from local warehouse $j$ to $k$ is $t_{j,k}^{\text{plt}} \geq t_{j,k}^{\text{rlt}}$ for $j, k \in J_l, j \neq k$. The final cost factor for transshipments that is used in this model are costs for performing such a proactive lateral transshipment, denoted by $c_{j,k}^{\text{plt}} \leq c_{j,k}^{\text{rlt}}$. Figure 3.1 depicts an overview of the transshipment flow of this network, where local warehouse $j$ is a main.

![Diagram of transshipment flow](image)

Figure 3.1: Transshipment time and cost flow

Decisions for allocation and proactive lateral transshipments are often made at fixed points in time rather than being made continuously. Therefore, time is discretized in time steps of length $\Delta t$. The discrete points in time mark the decision epochs at which the information about events in the previous time step becomes available to the decision maker. We can then define $P_{i,j}(\Delta t)$ as the probability that a part of SKU $i$ at local warehouse $j$ has failed during discrete time step $\Delta t$. For the ease of notation, and without loss of generality, $\Delta t = 1$ is chosen and the dependence on $\Delta t$ is dropped. Then, since failures arrive according to a Poisson process with rate $\lambda_{i,j}$, we have

$$P_{i,j} := P(Z_{i,j} < 1) = 1 - e^{-\lambda_{i,j}},$$  \hspace{1cm} (3.1)$$

where $Z_{i,j}$ denotes the lifetime distribution for SKU $i$ at local warehouse $j$, which is an exponentially distributed random variable.
3.2 Service level agreements

A common practice in spare parts inventory systems, is the use of service level agreements (SLAs). These ensure a certain service performance at the end of the contract period. Let $T$ denote the duration of a contract period. In our model, each local warehouse has its own SLA. Penalty cost may be incurred if these SLAs are not met. The costs for not meeting the agreements are often very high, which makes it crucial to manage the spare parts system such that the SLAs are met against minimal costs.

To ensure that the SLAs are met from tactical planning perspective, spare parts are put on stock in each local warehouse, called basestock levels. Let $S_{i,j}$ denote the basestock level for SKU $i \in I$ in local warehouse $j \in J$. Furthermore, $S_i := (S_{i,1}, \ldots, S_{i,|J|})$ denotes the vector of basestock levels for SKU $i \in I$, and let the basestock levels for the whole system be denoted by

$$
S = 
\begin{pmatrix}
S_{1,1} & S_{1,2} & \cdots & S_{1,|J|} \\
S_{2,1} & S_{2,2} & \cdots & S_{2,|J|} \\
\vdots & \vdots & \ddots & \vdots \\
S_{|I|,1} & S_{|I|,2} & \cdots & S_{|I|,|J|}
\end{pmatrix}.
$$

We assume that these basestock levels are already set and fixed by the planning method of Van Houtum and Kranenburg (2015, Ch. 5) on a tactical level. Although their model makes basestock level decisions for only one region, basestock levels for the entire network can be determined by calculating the basestock levels for each region separately. A similar approach is taken by Van Aspert (2014). The holding costs form a constant factor, and therefore they are excluded from the model.

For a spare parts system, failure rates are often very low, which leads to low basestock levels. ASML confirms that this is true in their inventory systems, and therefore we assume that all basestock levels in our model are at most two. As each part failure triggers a new part from the supplier, the number of parts in the system is constant for each SKU $i$. Let $OH_{i,j} \in \mathbb{N}_0$ denote the on-hand inventory for SKU $i$ at local warehouse $j$. For local warehouses also the parts in the pipeline are counted, such that $OH_{i,j}$ is equal to the actual on-hand stock plus parts that are in the pipeline. Parts in the pipeline are either parts of SKU $i$ that are allocated to local warehouse $j$ or parts that are transshipped to local warehouse $j$ proactively. As stated before, the global warehouse does not hold stock, which implies that all parts that are not on-hand at a local warehouse, are in repair or production. Since the supplier has ample capacity, the number of deriving parts is the sum of $\sum_j (S_{i,j} - OH_{i,j})$ Bernoulli trials. Here, the probability of success stands for a part arriving at the global warehouse, and is equal to $1/t_i^*$. Hence, for $n = \sum_j (S_{i,j} - OH_{i,j})$, the probability of $k$ parts arriving at the global warehouse during one time step is

$$p_i^{GW}(k) := \binom{n}{k} \left( \frac{1}{t_i^*} \right)^k \left( 1 - \frac{1}{t_i^*} \right)^{n-k}.$$  \hspace{1cm} (3.2)
These $n$ independent Bernoulli experiments capture part of the randomness due to supplier unreliability while keeping the problem tractable through its memoryless property.

The service level agreements in this model are defined by a maximum number of allowed stockouts for the entire contract period. This target is defined for each local warehouse individually and is denoted by $\tau_j(T)$ for $j \in J_l$. This $\tau_j(T)$ is based on the basestock levels, the demand and service rate of the SKUs. If, at the end of a service period this target is not met, a fixed penalty cost per local warehouse will be incurred, which is denoted by $c_{j \text{pen}}$. Each failure during the contract period with $t$ time left decreases $\tau_j(t)$ by 1, such that $\tau_j(t)$ is non-decreasing in $t$, i.e. $\tau_j(T) \geq \tau_j(T-1) \geq \cdots \geq \tau_j(t) \geq \cdots \geq \tau_j(1) \geq \tau_j(0)$.

### 3.3 Overview of Assumptions

Several assumptions that underlie this model are introduced in the previous sections. This section presents and briefly discusses these assumptions:

- **Parts criticality**
  All spare parts are equally critical for the functioning of equipments. That is, each part request corresponds to a failure that causes machine down.

- **Perfect diagnosis**
  Upon a failure of a system, a perfect diagnosis can be executed to determine which part caused the failure.

- **Perfect parts**
  Parts of the same type arriving at the global warehouse are perfect and equal, i.e. they always get the machine back up and running. This also implies that a so-called dead on arrival will not occur.

- **Unit demand**
  Each failure corresponds to a request of only one unit of spare part from stock.

- **One-for-one replenishment**
  An one-for-one replenishment policy $(S-1, S)$ is applied for the replenishment of spare parts stock at the local warehouses. That is, a replenishment order, either repair or production, is placed for each spare part requested from the local warehouse. All SKUs are subject to the one-for-one replenishment strategy.

- **Independent replenishments**
  The replenishment lead time from the global warehouse to the local warehouse will
be different for each part. Replenishment lead times for different SKUs are independent, and replenishment lead times for the same SKU are independent and identically distributed (i.i.d).

- **Ample capacity**
  The supplier has ample capacity for repairing defective parts and producing new parts. This means that there are no parts waiting to start repair or production and finished parts are sent to the global warehouse independent of the repair or production process of other parts.

- **Poisson distribution**
  Each part fails according to a Poisson distribution with a fixed mean that is equal to the failure rate of the part. The demand processes for all items at all locations are mutually independent Poisson processes with known failure rates, i.e. part failures happen independent of each other. Failure of a part does not trigger, in a positive or negative way, failures of other parts in the same equipment. Furthermore, the interarrival times of failures for a particular SKU at a particular local warehouse are i.i.d.

- **Local supply**
  Each customer is supplied from a specific predefined local warehouse. The transportation time and cost between the local warehouse and the customer is negligible.

- **Lateral supply**
  If a local warehouse is out-of-stock then the requested part can be supplied from other local warehouses in a known predefined order. The lateral transshipment times and costs between local warehouses are fixed and known.

- **Fixed basestock levels**
  The basestock levels that are calculated at a tactical level are assumed to be fixed through time, so no changes are made on tactical level during the contract period. Moreover, these basestock levels are assumed to be less than or equal to two.
Chapter 4

Heuristic Approach

In this chapter, a heuristic approach is developed for deciding on proactive lateral transshipments and stock allocations. This approach is based on the transient behavior of Erlang loss systems as mentioned in Section 2.5. First, the distribution for the number of stockouts is presented, which will be used to calculate the probability of exceeding the allowed number of stockouts until the end of the contract period for each warehouse. Based on these probabilities, a cost function for the expected costs until the end of the contract period is derived. We formulate two optimization problems that minimize this cost function. The first problem only makes stock allocation decisions, and is used for comparison with other allocation-only policies. The second problem also allows for proactive lateral transshipments between local warehouses, to investigate their added value.

In Chapter 5 and Chapter 6 the performance of FCFS, the current allocation policy of ASML and the two proposed heuristics is compared. Here, the model characteristics are as described in Chapter 4, which means that there are fixed rules for reactive lateral transshipments and emergency shipments. All possible transshipments are made at the beginning of each time step. That means reactive lateral transshipments and emergency shipments as well as allocation shipments, if parts are available at the global warehouse, and proactive lateral transshipments, if this will benefit the performance of the entire network. For the proposed heuristics, the transshipments that minimize the two minimization problems are the allocation shipments and proactive transshipment that should be made in this time step.

4.1 Isolated Warehouses

For the two heuristics, we assume that all local warehouses operate individually, which means that we neglect the occurrence of any reactive lateral transshipments or emergency shipments in case of a local warehouse experiencing a stockout. The rate at which failures occur is quite low, such that the number of reactive lateral transshipments in a finite period does not have a significant influence on the transient behavior of the individual loss systems. These low failures rates are also confirmed by ASML. Moreover, including these reactive transshipments would extremely complicate the analysis.
of these loss systems, with such interacting local warehouses.

Throughout this chapter, the characteristics of Erlang loss systems are investigated. Therefore, we translate the spare parts model to a model considered in queueing theory, including the concepts and notation used. Each local warehouse is modeled as an Erlang loss queueing system. For that, the basestock levels correspond to the number of servers, such that the on-hand inventory level corresponds to the number of idle servers. Hence, the number of busy servers corresponds to the number of parts being replenished, i.e. in repair or production. The number of part requests due to failures corresponds to customer arrivals, and the repair and production completions are the service completions. Furthermore, an Erlang loss system has no waiting room, and customers arriving when all servers are occupied, are lost. Hence, the word losses is used rather than stockouts in a queueing model. As the inventory and queueing models coincide from a mathematical perspective, the results and insights for the queueing problems can be translated into results and insights for spare parts problems. For all individual warehouses, the failure and replenishment process for each SKU \(i\) at local warehouse \(j\) can thus be modeled as a \(M/M/S_{i,j}/S_{i,j}\) queueing system. Therefore, probability of losses are basestock level specific, i.e. these probabilities are different for base stock levels 0, 1, 2. In the next section, the distribution of the number of losses is derived for each of these basestock levels.

### 4.2 Distribution of the Number of Losses

In this section, we present a procedure to find expressions of the time-dependent probability distribution of the number of losses. This is dependent on the total number of servers, the number of busy servers at time 0, the length of the interval \((0, t)\), and the arrival and service rates. In our model, only basestock levels less than or equal to 2 are considered, as mentioned in Section 3.2. Therefore, the \(M/M/0/0\), \(M/M/1/1\) and \(M/M/2/2\) loss systems will be analyzed in Section 4.2.1 – 4.2.3 respectively. A general approach to find the distribution of the number of losses is provided, such that this can also be applied for a \(M/M/K/K\) loss system with \(K > 2\), as provided in Section 4.2.4. As confirmed by interviews at ASML, the probability of more than two losses during the contract period is significantly low. Therefore, expressions for these probabilities will not be given, and only the probability of zero, one or two losses are considered. In the following sections, all calculations and the expressions for the probability of losses hold for general arrival rate \(\lambda\), service rate \(\mu\) and time \(t\).

#### 4.2.1 \(M/M/0/0\) System (Basestock Level 0)

For items that have basestock level 0, there is never stock at the local warehouse. No stock in the local warehouse translates to a queueing system with zero servers and zero waiting room.
Therefore, the number of losses simply reduces to the number of arrivals (failures) for that part. The probability of \( n \in \mathbb{N}_0 \) losses in an interval of length \( t \) is denoted by

\[
\mathbb{P}(L(t) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!},
\]

where \( L(t) \) denotes the number of losses in interval \((0, t)\).

### 4.2.2 M/M/1/1 System (Basestock Level 1)

Similar to Ferrante (2009), the probability of \( n \) losses, \( (n \in \mathbb{N}_0) \), in interval \((0, t)\) for a M/M/1/1 queueing system needs to be found. Let \( f_x(n, t) = \mathbb{P}(L(t) = n|X(0) = x) \) for \( x \in \{0, 1\} \), where \( X(t) \) denotes the number of busy servers at time \( t \). Then,

\[
\begin{align*}
    f_0(n, t) &= e^{-\lambda t} \mathbb{1}_{\{n=0\}} + \int_{u=0}^{t} \lambda e^{-\lambda u} f_1(n, t-u) \, du, \\
    f_1(n, t) &= e^{-\mu t} e^{-\lambda t} \frac{(\lambda t)^n}{n!} + \int_{u=0}^{t} \mu e^{-\mu u} \sum_{m=0}^{n} \frac{e^{-\lambda u} (\lambda u)^m}{m!} f_0(n-m, t-u) \, du.
\end{align*}
\]

This expression for an initial situation with zero busy servers, \( f_0(n, t) \), is obtained by conditioning on the first arrival. No arrival in \((0, t)\) would only be possible in case no losses are required, i.e. \( n = 0 \) (the first term). If a customer does arrive, say at time \( u \in (0, t) \), it positions us in a situation with one busy server and \( n \) losses to occur over an interval of length \( t - u \), for \( n \geq 0 \) (the second term). In case there is one busy server initially, \( f_0(n, t) \) is obtained by conditioning on the first service completion. If the first service is not completed in \((0, t)\), then all \( n \geq 0 \) losses must occur before time \( t \) (the first term). If a service is completed at time \( u \in (0, t) \), we sum over all possibilities of having already \( m, 0 \leq m \leq n \) losses in time \((0, u)\). The remaining \( n-m \) losses should still occur in the remaining \( t-u \) time for a situation with zero busy servers (the second term).

Convolutions as in the system of equations in (4.1) are often transformed using the Laplace transform. Using the Laplace transform given by \( \phi_{\lambda}(n, s) = \int_{t=0}^{\infty} e^{-st} f_x(n, t) \, dt \) for \( x \in \{0, 1\} \), we find

\[
\begin{align*}
    \phi_0(n, s) &= \int_{t=0}^{\infty} e^{-st} e^{-\lambda t} \mathbb{1}_{\{n=0\}} \, dt + \int_{t=0}^{\infty} e^{-st} \int_{u=0}^{t} \lambda e^{-\lambda u} f_1(n, t-u) \, du \, dt, \\
    \phi_1(n, s) &= \int_{t=0}^{\infty} e^{-st} e^{-\mu t} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \, dt \\
    &\quad+ \int_{t=0}^{\infty} e^{-st} \int_{u=0}^{t} \mu e^{-\mu u} \sum_{m=0}^{n} \frac{e^{-\lambda u} (\lambda u)^m}{m!} f_0(n-m, t-u) \, du \, dt.
\end{align*}
\]

For \( \phi_1(n, s) \) it is used that for the gamma function \( \Gamma(y) = \int_{t=0}^{\infty} t^{y-1} e^{-t} \, dt \) it holds that \( \Gamma(y) = (y-1)! \).
for \( y \in \mathbb{N}^+ \), and by partial integration, we know that
\[
\int_{t=0}^{\infty} \lambda e^{-\lambda t} \frac{(\lambda t)^n}{n!} \, dt = \int_{t=0}^{\infty} (s + \mu + \lambda) e^{-st} e^{-\mu t} \frac{(s + \mu + \lambda) t^n}{n!} \, dt = 1.
\]

This can then be used for the first part of \( \phi_1(n, s) \) in Equation (4.2), such that
\[
\int_{t=0}^{\infty} e^{-st} e^{-\mu t} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \, dt = \left( \frac{\lambda}{\lambda + \mu + s} \right)^n \frac{1}{\lambda + \mu + s}.
\] (4.3)

For the second part, using the convolution and interchanging the order of integration leads to
\[
\int_{t=0}^{\infty} e^{-st} \int_{u=0}^{t} \mu e^{-\mu u} \sum_{m=0}^{n} e^{-\lambda u} \frac{(\lambda u)^m}{m!} f_0(n - m, t - u) \, du \, dt
\]
\[
= \sum_{m=0}^{n} \phi_0(n - m, s) \frac{\mu}{\lambda + \mu + s} \left( \frac{\lambda}{\lambda + \mu + s} \right)^m.
\] (4.4)

Then, the Laplace transform of \( f_0(n, t) \) can be analyzed, and Equation (4.3) and (4.4) can be substituted in (4.2), which results in
\[
\phi_0(n, s) = \frac{1}{\lambda + s} \mathbb{1}_{\{n=0\}} + \frac{\lambda}{\lambda + s} \phi_1(n, s),
\]
\[
\phi_1(n, s) = \left( \frac{\lambda}{\lambda + \mu + s} \right)^n \frac{1}{\lambda + \mu + s} + \sum_{m=0}^{n} \phi_0(n - m, s) \frac{\mu}{\lambda + \mu + s} \left( \frac{\lambda}{\lambda + \mu + s} \right)^m.
\]

Now, we make use of generating function of the Laplace transform, i.e. \( F_x(z, s) = \sum_{n=0}^{\infty} \phi_x(n, s) z^n \), and find a set of equations for \( F_0(z, s) \) and \( F_1(z, s) \), which is given by
\[
F_0(z, s) = \frac{1}{\lambda + s} + \frac{\lambda}{\lambda + s} F_1(z, s),
\]
\[
F_1(z, s) = \frac{1}{\lambda + \mu + s} \frac{1}{1 - \frac{\lambda z}{\lambda + \mu + s}} + \frac{1}{1 - \frac{\lambda z}{\lambda + \mu + s}} F_0(z, s) \frac{\mu}{\lambda + \mu + s}.
\]

Solving this set of equations, we find for \( F_0(z, s) \) and \( F_1(z, s) \):
\[
F_0(z, s) = \frac{\lambda(2 - z) + \mu + s}{s^2 + s(\lambda(2 - z) + \mu) + \lambda^2(1 - z)},
\]
\[
F_1(z, s) = \frac{\lambda + \mu + s}{s^2 + s(\lambda(2 - z) + \mu) + \lambda^2(1 - z)}.
\]

From here, taking the inverse Laplace transform of \( F_0(z, s) \) and \( F_1(z, s) \) results in the probability generating function of the distribution of the number of losses. Partial fraction decomposition is used to enable partwise inversion of the Laplace transform of the generating function. Using partial
fraction decomposition, $F_0(z, s)$ and $F_1(z, s)$ can be written as

$$F_0(z, s) = \frac{\lambda(2 - z) + \mu + s}{s^2 + s(\lambda(2 - z) + \mu) + \lambda^2(1 - z)} = \frac{A_{0,1}(z)}{s - s_1(z)} + \frac{A_{0,2}(z)}{s - s_2(z)},$$

$$F_1(z, s) = \frac{\lambda + \mu + s}{s^2 + s(\lambda(2 - z) + \mu) + \lambda^2(1 - z)} = \frac{A_{1,1}(z)}{s - s_1(z)} + \frac{A_{1,2}(z)}{s - s_2(z)},$$

where the roots of the denominator are given by

$$s_1(z), s_2(z) = \frac{-2\lambda - \mu + \lambda z \pm \sqrt{4\lambda \mu + (\mu - \lambda z)^2}}{2},$$

and the corresponding terms for $F_0(z, s)$ and $F_1(z, s)$ in the numerator of the partial fraction decomposition are

$$A_{0,1}(z), A_{0,2}(z) = \frac{1}{2} \pm \frac{2\lambda + \mu - \lambda z}{2\sqrt{4\lambda \mu + (\mu - \lambda z)^2}},$$

$$A_{1,1}(z), A_{1,2}(z) = \frac{1}{2} \pm \frac{\mu + \lambda z}{2\sqrt{4\lambda \mu + (\mu - \lambda z)^2}}.$$

Then, the inverse Laplace transform of $\phi_x(n, s)$ can be written as

$$H_x(z, t) = \sum_{n=0}^{\infty} \mathbb{P}(L(t) = n|X(0) = x)z^n = A_{x1}(z)e^{s_1(z)t} + A_{x2}(z)e^{s_2(z)t}.$$

This leads to the probability generating functions

$$H_0(z, t) = \left( \frac{1}{2} + \frac{2\lambda + \mu - \lambda z}{2\sqrt{4\lambda \mu + (\mu - \lambda z)^2}} \right) e^{-\frac{1}{2}(2\lambda + \mu - \lambda z - \sqrt{4\lambda \mu + (\mu - \lambda z)^2})t} + \left( \frac{1}{2} - \frac{2\lambda + \mu - \lambda z}{2\sqrt{4\lambda \mu + (\mu - \lambda z)^2}} \right) e^{-\frac{1}{2}(2\lambda + \mu - \lambda z + \sqrt{4\lambda \mu + (\mu - \lambda z)^2})t},$$

$$H_1(z, t) = \left( \frac{1}{2} + \frac{\mu + \lambda z}{2\sqrt{4\lambda \mu + (\mu - \lambda z)^2}} \right) e^{-\frac{1}{2}(2\lambda + \mu - \lambda z - \sqrt{4\lambda \mu + (\mu - \lambda z)^2})t} + \left( \frac{1}{2} - \frac{\mu + \lambda z}{2\sqrt{4\lambda \mu + (\mu - \lambda z)^2}} \right) e^{-\frac{1}{2}(2\lambda + \mu - \lambda z + \sqrt{4\lambda \mu + (\mu - \lambda z)^2})t}.$$

Then, the exact expressions for the number of losses given the current number of busy servers $X(0)$ can be found by using

$$\mathbb{P}(L(t) = n|X(0) = x) = f_x(n, t) = \left( \frac{1}{n!} \right) \frac{d^n}{dz^n} H_x(z, t) \bigg|_{z=0}. \quad (4.5)$$

The expressions for $\mathbb{P}(L(t) = n|X(0) = x)$ for $n \in \{0, 1, 2\}$ and $x \in \{0, 1\}$ after applying Equation (4.5) can be found in Appendix A. We note that for $z = 1$, the generating functions $H_0(z, t)$ and
4.2 Distribution of the number of losses

$H_1(z, t)$ are equal to the sum of the probabilities and should be equal to 1. Indeed, $H_0(1, t) = 1$ and $H_1(1, t) = 1 \forall t$. Furthermore, for $z = 0$ and $t \to \infty$, $H_0(z, t)$ and $H_1(z, t)$ represent the probability of zero losses in $(0, \infty)$ given zero and one busy servers at time 0, respectively. We find that, for $\lambda > 0$, $H_0(0, t) \xrightarrow{t \to \infty} 0$ and $H_1(0, t) \xrightarrow{t \to \infty} 0$. This is logical, as there will be at least one loss in an infinity amount of time. For $\lambda = 0$, both limits tend to 1, as with no demand there will also be no losses.

4.2.3 $M/M/2/2$ system (Basestock level 2)

For the $M/M/2/2$ queueing systems, the same steps can be taken to find the distribution for the number of losses as was done for the $M/M/1/1$ queueing system. Hence,

\[
\begin{align*}
 f_0(n, t) &= e^{-\lambda t} \mathbb{1}_{\{n=0\}} + \int_{u=0}^{t} \lambda e^{-\lambda u} f_1(n, t-u) \, du, \\
 f_1(n, t) &= e^{-(\lambda + \mu) t} \mathbb{1}_{\{n=0\}} + \int_{u=0}^{t} (\lambda + \mu) e^{-(\lambda + \mu) u} \left( \frac{\mu}{\lambda + \mu} f_0(n, t-u) + \frac{\lambda}{\lambda + \mu} f_2(n, t-u) \right) \, du, \\
 f_2(n, t) &= e^{-2\mu t} e^{-\lambda t} \frac{(\lambda t)^n}{n!} + \int_{u=0}^{t} 2\mu e^{-2\mu u} \sum_{m=0}^{n} e^{-\lambda u} \left( \frac{\mu}{\lambda + \mu} \right)^m f_1(n-m, t-u) \, du.
\end{align*}
\]

(4.6)

Here, $f_0(n, t)$ and $f_2(n, t)$ are achieved in a similar analysis as in Equation (4.1). For $f_1(n, t)$, this analysis is slightly different. For an initial situation of one busy server, again $n$ losses can be achieved if no customer arrivals and no service is completed in case $n = 0$. There are two other options for transitions between states. On the one hand we can shift towards a situation with two busy servers that still requires $n$ losses in a time interval of length $t-u$ if at time $u$ a customer arrives before a service is completed. On the other hand, if a service is completed at time $u$ before a customer arrives, there are now zero busy servers and there are still $n$ losses left in the remaining period of length $t-u$.

In order to deal with the convolution in Equation (4.6), we again resort to Laplace transforms ($\phi_x(n, s) = \int_{t=0}^{\infty} e^{-st} f_x(n, t) \, dt$ for $x \in \{0, 1, 2\}$) and find

\[
\begin{align*}
 \phi_0(n, s) &= \frac{1}{\lambda + s} \left( \mathbb{1}_{\{n=0\}} + \lambda \phi_1(n, s) \right), \\
 \phi_1(n, s) &= \frac{1}{\lambda + \mu + s} \left( \mathbb{1}_{\{n=0\}} + \mu \phi_0(n, s) + \lambda \phi_2(n, s) \right), \\
 \phi_2(n, s) &= \frac{1}{\lambda + 2\mu + s} \left( \left( \frac{\lambda}{\lambda + 2\mu + s} \right)^n + 2\mu \sum_{m=0}^{n} \phi_1(n-m, s) \left( \frac{\lambda}{\lambda + 2\mu + s} \right)^m \right).
\end{align*}
\]
Hereafter, the Laplace generating functions, i.e. \( F_x(z,s) = \sum_{n=0}^{\infty} \phi_x(n,s)z^n \), are used to find

\[
\begin{align*}
F_0(z,s) &= \frac{1}{\lambda + s}(1 + \lambda F_1(z,s)), \\
F_1(z,s) &= \frac{1}{\lambda + \mu + s}(1 + \mu F_0(z,s) + \lambda F_2(z,s)), \\
F_2(z,s) &= \frac{1}{\lambda + 2\mu + s - \lambda z}(1 + 2\mu F_1(z,s)).
\end{align*}
\]

This set of equations can again be solved, which results in

\[
\begin{align*}
F_0(z,s) &= \frac{s^2 + (3\lambda + 3\mu - \lambda z)s - 2\lambda^2 z + 3\lambda^2 - \lambda \mu z + 3\lambda \mu + 2\mu^2}{s^3 + (3\mu + 3\lambda - \lambda z)s^2 + (3\lambda \mu - 2\lambda^2 z + 3\lambda^2 + 2\mu^2 - \lambda \mu z)s + \lambda^3 z + \lambda^3}, \\
F_1(z,s) &= \frac{s^2 + (3\lambda + 3\mu - \lambda z)s - \lambda^2 z + 2\lambda^2 - \lambda \mu z + 3\lambda \mu + 2\mu^2}{s^3 + (3\mu + 3\lambda - \lambda z)s^2 + (3\lambda \mu - 2\lambda^2 z + 3\lambda^2 + 2\mu^2 - \lambda \mu z)s + \lambda^3 z + \lambda^3}, \\
F_2(z,s) &= \frac{s^2 + (2\lambda + 3\mu)s + \lambda^2 + 2\lambda \mu + 2\mu^2}{s^3 + (3\mu + 3\lambda - \lambda z)s^2 + (3\lambda \mu - 2\lambda^2 z + 3\lambda^2 + 2\mu^2 - \lambda \mu z)s + \lambda^3 z + \lambda^3}.
\end{align*}
\]

These functions have complex roots in the denominator when looking for a partial fraction decomposition. Since the partial fraction decomposition cannot be found as easily as in the \( M/M/1/1 \) system, the Heaviside cover-up rule is used (Adams and Essex, 2010, p. 340).

Using the Heaviside cover-up rule, the denominator (of degree \( n \)) of \( F_i(z,s) \) can be written as a product of (distinct) linear factors (i.e. \( s_k \neq s_l \forall k, l \in \{1, \ldots, n\} \)). For \( n = 3 \) this reduces to

\[
F_x(z,s) = \frac{p_x(s)}{(s-s_1(z))(s-s_2(z))(s-s_3(z))} = \sum_{k=1}^{3} \frac{A_{x,k}(z)}{s-s_k(z)}, \tag{4.7}
\]

Then we can find \( A_{x,k}(z) \) by multiplying \( F_x(z,s) \) by the factor \( (s-s_k(z)) \) and substituting \( s = s_k(z) \) in the rest of the expression, giving

\[
A_{x,k}(z) = (s-s_k(z))F_x(z,s)|_{s=s_k(z)} \tag{4.8}
\]

With these \( A_{x,k}(z) \), the probability generating function is written as an exponential sum:

\[
H_x(z,t) = \sum_{n=0}^{\infty} P(L(t) = n | X(0) = x)z^n = A_{x,1}(z)e^{s_1(z)t} + A_{x,2}(z)e^{s_2(z)t} + A_{x,3}(z)e^{s_3(z)t}.
\]

It follows that

\[
H_x(z,t) = \frac{\sum_{k=1}^{3} e^{s_k(z)t}\left(s_{\nu_{k,3}}(z) - s_{\nu_{k+1,3}}(z)\right)g_{x,k}(z)}{(s_1(z) - s_2(z))(s_1(z) - s_3(z))(s_2(z) - s_3(z))},
\]
4.2 Distribution of the number of losses

where

\[
g_{x,k}(z) = \begin{cases} 
  s_k(z)^2 + (3\lambda + 3\mu - \lambda z)s_k(z) - 2\lambda^2 z + 3\lambda^2 - \lambda \mu z + 3\lambda \mu + 2\mu^2 & \text{for } x = 0, \\
  s_k(z)^2 + (3\lambda + 3\mu - \lambda z)s_k(z) - \lambda^2 z + 2\lambda^2 - \lambda \mu z + 3\lambda \mu + 2\mu^2 & \text{for } x = 1, \\
  s_k(z)^2 + (2\lambda + 3\mu)s_k(z) + \lambda^2 + 2\lambda \mu + 2\mu^2 & \text{for } x = 2,
\end{cases}
\]

and \( \nu_{k,x} \) is defined by \( \nu_{k,x} := k \mod x + 1 \). Furthermore, the denominator of \( H_x(z, t) \) factors by \( s^3 + (3\mu + 3\lambda - \lambda z)s^2 + (3\lambda \mu - 2\lambda^2 z + 3\lambda^2 + 2\mu^2 - \lambda \mu z)s + \lambda^3 z + \lambda^3 = (s - s_1(z))(s - s_2(z))(s - s_3(z)) \), where \( s_1(z) \), \( s_2(z) \) and \( s_3(z) \) are given by

\[
\begin{align*}
  s_1(z) &= \frac{1}{12} \left[ 2 \cdot \frac{2 \sqrt{3} T (\lambda^2 z^3 - 3\lambda \mu(z - 3) + 3\mu^2)}{\alpha(z)} + 2 \cdot \frac{\alpha(z)}{2^{-2/3}} - 12(\lambda + \mu) + 4\lambda z \right], \\
  s_2(z) &= \frac{1}{12} \left[ (-1 - \sqrt{3} i) \frac{2 \sqrt{3} T (\lambda^2 z^3 - 3\lambda \mu(z - 3) + 3\mu^2)}{\alpha(z)} + (1 + \sqrt{3} i) \frac{\alpha(z)}{2^{-2/3}} - 12(\lambda + \mu) + 4\lambda z \right], \\
  s_3(z) &= \frac{1}{12} \left[ (-1 + \sqrt{3} i) \frac{2 \sqrt{3} T (\lambda^2 z^3 - 3\lambda \mu(z - 3) + 3\mu^2)}{\alpha(z)} + (1 - \sqrt{3} i) \frac{\alpha(z)}{2^{-2/3}} - 12(\lambda + \mu) + 4\lambda z \right].
\end{align*}
\]

where

\[
\alpha(z) = \left[ \sqrt{\lambda^2 (2\lambda^2 z^3 - 9\lambda \mu z^2 + 9\mu^2(z - 3))^2 - 4 (\lambda^2 z^3 - 3\lambda \mu(z - 3) + 3\mu^2)^3} + 2\lambda^3 z^2 - 9\lambda^2 \mu z^2 + 9\lambda \mu^2(z - 3) \right]^{1/3}.
\]

Note that \( s_1(z) \) is real, but that both \( s_2(z) \) and \( s_3(z) \) in Equation (4.9) are complex and that they are complex conjugates of each other. However, \( H_x(z, t) \) is not complex, as is shown in Appendix B.

Then again, Equation (4.5) is used for the expressions for \( P(L(t) = n|X(t) = x) \). If we then fill in \( z = 0 \) we find for \( P(L(t) = 0|X(0) = x) \) for \( x \in \{0, 1, 2\} \) the following:

\[
P(L(t) = 0|X(0) = x) = H_x(0, t) = \frac{\sum_{j=1}^{3} e^{s_j(0)t} \left(s_{\nu_{j,3}}(0) - s_{\nu_{j+1,3}}(0)\right) g_{x,j}(0)}{(s_1(0) - s_2(0))(s_3(0) - s_2(0))(s_1(0) - s_3(0))}.
\]

Furthermore, taking the first derivative of \( H_x(z, t) \) with respect to \( z \), denoted by \( ' \), gives

\[
\frac{d}{dz} H_x(z, t) = \sum_{k=1}^{3} \frac{e^{s_k(z)t}}{s_k(z) - s_{\nu_{k,3}}(z)} \left[ g_{x,k}(z) + s_k'(z)g_{x,k}(z) - \frac{g_{x,k}(z) s_k'(z) - s_{\nu_{k,3}}'(z)}{s_k(z) - s_{\nu_{k,3}}(z)} - \frac{g_{x,k}(z) \left(s_k'(z) - s_{\nu_{k+1,3}}'(z)\right)}{s_k(z) - s_{\nu_{k+1,3}}(z)} \right].
\]
Full expansions of these shorter-hand expressions are too complex and too long to write them down. Therefore, the general formula given by Equation (4.5) is used for any further calculations and implementation.

4.2.4 \( M/M/K/K \) system (Basestock level \( K \))

Although Erlang loss systems with more than 2 servers are not considered in this thesis, a general approach was found for finding the distribution of the number of losses in \((0, t)\). Based on the analysis for the \( M/M/1/1 \) and \( M/M/2/2 \) systems, several steps have to be taken to find this distribution. First, write down the probability of \( n \) losses with convolutions similar to Equation (4.1) and (4.6). Then consecutively use the Laplace transform \((\phi_x(n, s) = \int_0^\infty e^{-st} f_x(n, t) \, dt)\) and the Laplace generating function \((F_x(z, s) = \sum_{n=0}^\infty \phi_x(n, s) z^n)\) to find a set of equations for \( F_0(z, s) \), \( F_1(z, s), \ldots, F_K(z, s) \). Performing these first steps, we can find the set of equations of \( F_0(z, s) \), \( F_1(z, s), \ldots, F_K(z, s) \) as follows:

\[
f_x(n, t) \quad \text{for} \quad x \in \{0, 1, \ldots, K\} \quad \text{is given by}
\]

\[
f_x(n, t) = \begin{cases} 
  e^{-\lambda t} \mathbb{1}_{\{n=0\}} + \int_0^t \lambda e^{-\lambda u} f_1(n, t-u) \, du, & \text{for } x = 0, \\
  e^{-\lambda x} \mathbb{1}_{\{n=0\}} + \int_0^t (\lambda + x \mu) e^{-(\lambda + x \mu) u} \left( \frac{x \mu}{\lambda + x \mu} f_{x-1}(n, t-u) + \frac{\lambda}{\lambda + x \mu} f_{x+1}(n, t-u) \right) \, du, & \text{for } x = 1, \ldots, K-1, \\
  e^{-\kappa \mu t} e^{-\lambda t} \frac{\lambda^n}{n!} + \int_0^t K \mu e^{-K \mu u} \sum_{m=0}^n \frac{\lambda^m u^m}{m!} f_{K-1}(n - m, t-u) \, du & \text{for } x = K.
\end{cases}
\]

Then, after using the Laplace transform, we obtain

\[
\phi_x(n, s) = \begin{cases} 
  \frac{1}{\lambda + s} (\mathbb{1}_{\{n=0\}} + \lambda \phi_1(n, s)) & \text{for } x = 0, \\
  \frac{1}{\lambda + x \mu + s} (\mathbb{1}_{\{n=0\}} + x \mu \phi_{x-1}(n, s) + \lambda \phi_{x+1}(n, s)) & \text{for } x = 1, \ldots, K-1, \\
  \frac{1}{\lambda + K \mu + s} \left( \left( \frac{\lambda}{\lambda + K \mu + s} \right)^n + K \mu \sum_{m=0}^n \phi_{K-1}(n - m, s) \left( \frac{\lambda}{\lambda + K \mu + s} \right)^m \right) & \text{for } x = K,
\end{cases}
\]

which, after using the Laplace generating function, leads to

\[
F_x(z, s) = \begin{cases} 
  \frac{1}{\lambda + s} (1 + \lambda F_1(z, s)) & \text{for } x = 0, \\
  \frac{1}{\lambda + j \mu + s} (1 + x \mu F_{x-1}(z, s) + \lambda F_{x+1}(z, s)) & \text{for } x = 1, \ldots, K-1, \\
  \frac{1}{\lambda + K \mu + s - \lambda z} (1 + K \mu F_{K-1}(z, s)) & \text{for } x = K,
\end{cases}
\]
This is a set of $K + 1$ equations with $K + 1$ unknowns, which can be solved. This results in expressions for $F_{x}(z,s)$ that all have a polynomial of order $K + 1$ in the denominator. A partial fraction decomposition can be found using the Heaviside cover-up rule given by Equation (4.7) and (4.8). This can again be translated into the probability generation function using the inverse Laplace transform, which in its turn can be used to find the distribution of the number of losses by using Equation (4.5).

4.3 Probability of exceeding allowed stockouts

Based on the probabilities given in the previous section, the probability of exceeding the number of allowed stockouts can be calculated. Combining all probabilities on zero, one or two losses for each part, all combinations that lead to exceeding the allowed number of stockouts can be found and used to calculate the probability of violating a contract. Let $L_{i,j}(t)$ be the number of losses in $(0, t)$ for item $i \in I$ at local warehouse $j \in J$ given the basestock level and the current on-hand inventory, i.e. $L_{i,j}(t) := L(t)|_{\lambda=\lambda_{i,j},\mu=\mu_{i,j}}$. The total number of losses in $(0, t)$ for local warehouse $j$ is defined by

$$L_j(t) := \sum_{i \in I} L_{i,j}(t).$$

Note that the appropriate loss probability must be chosen from the previous section, based on the current on-hand inventory and the basestock level. Then the probability of exceeding the number of allowed stockouts for local warehouse $j$ can be denoted by

$$P(L_j(t) > \tau_j(t)) = P\left(\sum_{i \in I} L_{i,j}(t) > \tau_j(t)\right) = 1 - \sum_{x=0}^{\tau_j(t)} P\left(\sum_{i \in I} L_{i,j}(t) = x\right),$$

where $\tau_j(t)$ denotes the number of remaining allowed stockouts for local warehouse $j$, as introduced in Section 3.2, with $t$ time steps left in the contract period. Recall that $c_{j}^{\text{pen}}$ denotes the fixed penalty cost for warehouse $j \in J$ if the contract is not met. For having $OH_j = (OH_{i,j})_{i \in I}$ inventory on-hand, the expected penalty costs for a warehouse $j$ are then given by

$$C(OH_j, t) := c_{j}^{\text{pen}} \cdot P(L_j(t) > \tau_j(t)). \tag{4.10}$$

Regarding the complexity of calculating the probability that $L_j(t)$ exceeds $\tau_j(t)$, we first look at the number of ways to have $x$ losses divided between $|I|$ parts, where each part has less than 3 losses, $C_{|I|}(x, 3)$. This can be determined by the restricted compositions of natural numbers.\(^1\)

which is given by
\[ C_{|I|}(x, 3) = \sum_{n=0}^{\lfloor x/3 \rfloor} (-1)^n \binom{|I|}{n} \left( \frac{|I| + x - 3n - 1}{|I| - 1} \right) \]

For example, for three parts and four allowed stockouts with a maximum of two stockouts per parts, the number of combinations is 6, i.e. (2,2,0), (2,0,2), (2,1,1), (1,2,1), (1,1,2), and (0,2,2). Indeed \( C_3(4, 3) = 6 \). For each of these 6 combinations, the probability of happening can be determined, which can then be combined with the options for less than four stockouts to find the probability of contract violation.

For low \(|I|\), this number of combinations remains relatively low. However, for increasing \(|I|\), the number of ways to divide \( x \) losses over \(|I|\) parts increases significantly. Then, another approach is to find an approximation of the distribution of \( L_j(t) \). This could be done in several ways, for example by using the central limit theorem. Here, several parts can be grouped based on their distribution of the number of losses, and each group can then be approximated by a Poisson distribution. One can interpret this Poisson distribution as an approximation for the Binomial distribution which in its turn can be used to find an approximation for the Normal distribution. Then a summation over several groups will lead to a Normal distribution again. Using these properties, one can focus on the group(s) of parts that will be the determining factor for the number of losses. This reduces the system to a system with less SKUs, which may be less complex computational-wise. Another approach towards the distribution of \( L_j(t) \) is the two-moment fit of all \( L_{i,j} \) to find an approximation of the distribution of \( L_j(t) \). Several methods are discussed in Van Houtum and Kranenburg (2015, Ch. 6), among which a two-moment fit by a negative binomial distribution, a mixture of Erlangian distributions and a hyperexponential distribution with balanced means.

### 4.4 Optimization Problem

As stated in the introduction of this chapter, two minimization problems will be formulated, where the first one only includes allocation decisions, and the second includes also proactive lateral transshipments. We start by introducing notation that will cover both models, and we will differentiate between the two problems by imposing different, problem specific constraints.

There are two types of shipment decisions: allocations from the global warehouse and proactive lateral transshipments. Let the size of these shipments be denoted by \( y = (y_{i,j,k})_{i \in I, j \in J_l, k \in J_l} \), where \( y_{i,j,k} \in \mathbb{N}_0 \) denotes the shipment size for item \( i \) from warehouse \( j \) to local warehouse \( k \in J_l \). These are either replenishments from CW, i.e. \( j = 0 \), or proactive lateral transshipments, i.e. \( j \in J_l \). Note that we want all parts to be allocated to the local warehouses such that the global warehouse cannot have stock on hand after these shipment decisions are made. Hence, shipments cannot be performed towards the global warehouse such that for \( y_{i,j,k} \) we have that \( k \neq 0 \). The costs
associated with all of these transshipments are denoted by $c^{tr}$ per transshipment. This is regardless of the distance between warehouses, the items that are sent, and whether they are allocation or proactive transshipments, i.e. $c^{tr} = c_{j,k}^{pl} = c_j^l \forall j, k \in J_l$. Although the global warehouse is forced to make these shipments costs as it cannot hold stock, this should not influence the transshipment decisions.

Let the inventory before and after decisions at local warehouse $j \in J$ be denoted by $OH_j$ and $OH'_j$, respectively. Here, $OH_j = (OH_{1,j}, \ldots, OH_{|I|,j})$ and $OH'_j = (OH'_{1,j}, \ldots, OH'_{|I|,j})$, where the inventory level for SKU $i$ at local warehouse $j$ before decisions is denoted by $OH_{i,j}$ and $OH'_{i,j} = OH_{i,j} - \sum_{k \in J_l} y_{i,j,k} + \sum_{k \in J_l} y_{i,k,j}$ denotes the inventory levels after decisions. Here, $\sum_{k \in J_l} y_{i,j,k}$ denotes all outgoing stock and $\sum_{k \in J_l} y_{i,k,j}$ denotes all incoming stock for warehouse $j$. Hence, based on the current state $OH_j$, and all incoming and outgoing transshipments, we can find the new state and calculate its corresponding costs.

At each time step, decisions are made to minimize the total expected cost until the end of the contract. These decisions are the solutions to minimization problems that are formulated below, and they are solved at every time step as part of the simulation. These optimization problems can be formulated as a Mixed Integer Non-Linear Program (MINLP) of which the objective is to find the transshipments that minimize the total costs until the end of the contract after transshipments are made. That is, minimize the total expected cost with $t - 1$ time steps left, because decisions made with $t$ time steps left take effect in the next period. So we take decisions that are optimal from the upcoming period forward. As these two minimization problems take the current contract performance into consideration, the first (allocation only) approach will be called Performance Anticipating Stock Allocation, or in short PASA. The second heuristic is called PASA-P, where the letter ‘P’ denotes the inclusion of proactive lateral transshipment.

The allocation-only problem (PASA) is formulated as

\[
\text{(PASA)} \quad \arg\min_{y} \sum_{j \in J_l} C(OH'_{j,t-1}) + c^{tr} \cdot \sum_{i \in I} \sum_{j \in J_l} \sum_{k \in J_l} y_{i,j,k},
\]

subject to

\[
\sum_{k \in J_l} y_{i,0,k} = OH_{i,0}, \quad \forall i \in I, \quad (4.11)
\]

\[
y_{i,0,j} \leq S_{i,j} - OH_{i,j}, \quad \forall i \in I, j \in J_l, \quad (4.12)
\]

\[
y_{i,j,k} = 0, \quad \forall i \in I, j, k \in J_l, \quad (4.13)
\]

\[
y_{i,0,k} \geq 0, \quad \forall i \in I, k \in J_l, \quad (4.14)
\]

where $C(OH'_{j,t-1})$ is defined by Equation (4.10). Furthermore, Equation (4.11) ensures that all stock that becomes available in the global warehouse will be sent into the field, i.e. divided over the local warehouses that have the highest need for this part; Equation (4.13) ensures that local
warehouses cannot send parts, i.e. no proactive lateral transshipments; Equation (4.14) states that
shipments from the global warehouse cannot have a negative size; and finally, transshipments to
the local warehouses cannot be performed if this results in overstocking in a local warehouse, which
is imposed by Equation (4.12).

For problem (PASA-P), including proactive lateral transshipments as well, the minimization problem becomes

\[
\text{(PASA-P)} \quad \arg\min_y \sum_{j \in J_l} C(OH_{j,t}^t, t-1) + c_tr \cdot \sum_{i \in I} \sum_{j \in J_l} \sum_{k \in J_l} y_{i,j,k},
\]

subject to

\[
\sum_{k \in J_l} y_{i,0,k} = OH_{i,0}, \quad \forall i \in I, \quad (4.15)
\]

\[
\sum_{k \in J_l} y_{i,j,k} \leq OH_{i,j}, \quad \forall i \in I, j \in J_l, \quad (4.16)
\]

\[
\sum_{k \in J} y_{i,k,j} - \sum_{k \in J_l} y_{i,j,k} \leq S_{i,j} - OH_{i,j}, \quad \forall i \in I, j \in J_l, \quad (4.17)
\]

\[
y_{i,j,k} \geq 0, \quad \forall i \in I, j \in J, k \in J_l. \quad (4.18)
\]

Here, Equation (4.15) is equal to (4.11) in (PASA); furthermore, proactive lateral transshipments can now occur and Equation (4.16) ensures that a local warehouse cannot send more parts than they have on-hand; Equation (4.18) states that shipments cannot have a negative size, not only for transshipment from the global warehouse, but this also holds for local warehouses; and finally, transshipments cannot be performed if this results in overstocking in a local warehouse, which is imposed by Equation (4.17). Also parts being sent away for local warehouses are taken into account.

For PASA-P, there are at most \( \prod_{i \in I, j \in J} (S_{i,j} + 1) \) possible combinations for all \( y \), and this number increases exponentially for an increasing number of parts and/or local warehouses. This implies that solving these problems using exhaustive search is extremely time-intensive for large systems. Hence, we resort to heuristic methods for solving these minimization problems. The greedy algorithm is used to make the allocation decisions for both PASA and PASA-P. This algorithm is described formally in Algorithm 1.

For each SKU \( i \in I \), a part is allocated to the local warehouse that reduces the total cost of the system the most. This is done iteratively, so one part at a time, until the global warehouse has no (more) parts to allocate.

For the decisions on proactive lateral transshipments, a similar approach is used. We use a local search method for finding the ‘neighbor’ of the current state that minimizes the current expected costs the most. That is, for all combinations of parts and local warehouses, we look for a transshipment of one part to another local warehouse that reduces the total expected costs until the end of the contract. This is performed in an iterative manner, so one part at a time, until there is no
Algorithm 1  Greedy Allocation Algorithm

1: for all $i \in I$ do
2:     while $OH_{i,0} > 1$ do
3:        for all $j \in J_l$ : $OH_{i,j} < S_{i,j}$ do
4:            Calculate $\sum_{\ell \in J_l} C(OH'_{\ell,t}, t-1)$ as defined in Equation (4.10), where $OH'_{i,j} = OH_{i,j} + 1$
5:        end for
6:        $j^* \leftarrow \arg\min_{j \in J_l : OH_{i,j} < S_{i,j}} C(OH'_{j,t}, t-1)$
7:        $OH_{i,j^*} \leftarrow OH_{i,j^*} + 1$
8:        $OH_{i,0} \leftarrow OH_{i,0} - 1$
9:     end while
10: end for

transshipment that reduces the expected cost anymore. The formal description for these proactive lateral transshipment decisions is described in Algorithm 2.

Algorithm 2 Proactive Lateral Transshipment Algorithm

1: $m \leftarrow \infty$
2: while $m > 0$ do
3:     $m \leftarrow \sum_{\ell \in J_l} C(OH_{\ell,t}, t-1)$
4:     $\tilde{m} \leftarrow 0$
5:     for all $j \in J_l$ and $i \in I : OH_{i,j} > 1$ do
6:         for all $k \in J_l : OH_{i,k} < S_{i,k}$ do
7:             Calculate $m_{i,j,k} := \sum_{\ell \in J_l} C(OH'_{\ell,t}, t-1)$ as defined in Equation (4.10), where $OH'_{i,k} = OH_{i,k} + 1$ and $OH'_{i,j} = OH_{i,j} - 1$
8:             if $m_{i,j,k} < \tilde{m}$ then
9:                 $\tilde{i} \leftarrow i$
10:                $\tilde{j} \leftarrow j$
11:                $\tilde{k} \leftarrow k$
12:                $\tilde{m} \leftarrow m - m_{i,j,k}$
13:         end if
14:     end for
15: end for
16: $OH_{i,j} \leftarrow OH_{i,j} - 1$
17: $OH_{i,k} \leftarrow OH_{i,k} + 1$
18: $m \leftarrow \tilde{m}$
19: end while
In this chapter, we present different aspects of the simulation design. First, the heuristics that are used to evaluate the performance of the two proposed heuristics are discussed. Second, we demonstrate how the service level agreements are computed. Third, we explain the different steps and procedures of the simulation, and show how these are verified.

5.1 Heuristics

In this section, the heuristic methods that are included in the simulation are listed. Obviously, the performance of the heuristics PASA and PASA-P should be investigated and these two models are included in the computational study. Furthermore, the performance of these heuristics is tested against two models used in practice. The first is the well-known FCFS rule, which is very intuitive and easy to implement. As stated in the introduction, we will also compare the performance of PASA and PASA-P with the performance of the allocation heuristic currently used within ASML, called the Network Oriented Replenishment Automation (NORA) policy. NORA uses the steady state blocking probabilities to find the number of stockouts for the current on-hand inventory level and compares this to the number of stockouts based on the basestock level. The difference between these two expected stockouts is called the unexpected stockouts, and these are used for the allocation of stock. A more detailed explanation of the calculations and steps of the NORA heuristic is given in Section 5.1.1.

As both heuristics used in practice are for allocation decisions only, the performance of these models will be compared with the PASA heuristic to see how much PASA can further improve the performance compared to how much the performance is already improved from using NORA rather than FCFS. Then, we will evaluate the added value of proactive lateral transshipments on top of the PASA heuristic.

These four heuristics are listed in Table 5.1, where it is indicated whether the current stock level, basestock level and contract performance are taken into account in the allocation decision process of each heuristic and whether proactive lateral transshipments can be made. The inclusion of the
contract performance includes taking into account both the remaining number of allowed stockouts and the time remaining until the end of the contract period.

Table 5.1: Overview of the four heuristics

<table>
<thead>
<tr>
<th>Name</th>
<th>Allocation</th>
<th>Proactive lateral transshipments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 FCFS</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2 NORA</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3 PASA</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>4 PASA-P</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

5.1.1 Network-Oriented Replenishments Automation

Within ASML, the NORA heuristic does more than only determining to which local warehouse stock should be allocated, such as creating repair orders, determining the stock needed for the installment of new machines and taking contracts into account where stock can come from anywhere in the region. In this thesis, we will only focus on the part that determines the decisions for allocation of stock to local warehouses that have a SLA in place for the number of parts that should be met from that local warehouse directly.

As mentioned above, allocations from the global warehouse are based on the unexpected number of stockouts, over a period of \( m \) time steps. Here, ‘unexpected’ is stated explicitly, as some stockouts are planned for while the basestock levels are calculated. So for example, parts that are very expensive are not always put on stock, but stockouts are taken into account in the calculation of the service level in the tactical planning, i.e. they are planned and compensated for by stocking more cheap parts such that the required performance is still met. Any additional risk for stockouts is captured in the unexpected number of stockouts. ASML uses the term unexpected non-availabilities or NAVs. It is calculated based on steady state performance of an Erlang loss system. The formula is given by

\[
\text{NAV}_{i,j} = [L(\text{OH}_{i,j}, t_i^s \cdot \lambda_{i,j}) - L(S_{i,j}, t_i^s \cdot \lambda_{i,j})] \cdot \lambda_{i,j} \cdot m, \tag{5.1}
\]

for part \( i \) at local warehouse \( j \), where \( \lambda_{i,j} \) is the failure rate per time step and \( L(c, \rho) \) denotes the Erlang Loss probability that is defined by

\[
L(c, \rho) = \frac{\rho^c / c!}{\sum_{x=0}^{c} \rho^x / x!},
\]

where \( c \) denotes the number of servers (in our case either the basestock level \( S_{i,j} \) or the current on-hand inventory level \( \text{OH}_{i,j} \) for part \( i \) at local warehouse \( j \)) and \( \rho \) denotes the offered workload (in our case the forecast demand during supply leadtime \( \lambda_{i,j} \cdot t_i^s \) for part \( i \) at local warehouse \( j \)). For example, if a basestock level of 3 is calculated with a forecast of 0.12 failure per time step
and a supply leadtime of 20 time steps and the current inventory level is 2, then the unexpected stockouts are \((L(2,2.4) - L(3,2.4)) \times (0.12 \times 365/12) = 0.69\).

If parts become available at the global warehouse due to a repair or production completion, NORA follows a multi-step approach to determine the best allocation decision, using these unexpected stockouts. This approach takes the following steps: as long as there is available supply, first the global warehouse is checked. In this warehouse, all parts are gathered in order to be allocated into the field, i.e. to the local warehouses. Furthermore, NORA states that in this global warehouse, a certain amount of each part should always be available at any time for emergencies. This amount is called the iron stock and is equal to one for all SKU. If stock is available in the global warehouse, it is stocked until iron stock level is met. So if possible, there is always one part of each SKU staying in the global warehouse. If parts are still available, the region with the highest unexpected stockouts is found. This is done by aggregating all demand, basestock levels and current inventory level for each region, and this is calculated by

\[
\text{NAV}_i^r = \left[ L \left( \sum_{j: r_j = r} O_{i,j} \cdot t_i^s \cdot \sum_{j: r_j = r} \lambda_{i,j} \right) - L \left( \sum_{j: r_j = r} S_{i,j} \cdot t_i^s \sum_{j: r_j = r} \lambda_{i,j} \right) \right] \cdot \sum_{j: r_j = r} \lambda_{i,j} \cdot m \tag{5.2}
\]

for region \(r\). After that, in the region with the highest NAV \((r^*)\), a part is allocated to the local warehouse with the highest NAV (= \(\text{argmin}_j \text{NAV}_{i,j}\) for \(r_j = r^*)\) as defined in Equation (5.1). Note that for both Equation (5.1) and (5.2), the variable \(m\) does not have added value since all terms are multiplied with \(m\) and it becomes a comparison between relative values. These allocation decisions are made one at a time, similar to Bertrand and Bookbinder (1998). So after one part is allocated and there is still stock available in the global warehouse, the complete procedure is repeated such that all NAVs are recalculated and an allocation is made based on these new NAVs. In Algorithm 3, this multi-step approach is described formally.

### 5.2 Allowed Number of Stockouts

As mentioned in Section 3.2, the basestock levels are determined such that the SLAs will be met. Currently, the basestock levels at ASML are set to meet the required performance based on steady-state performance and based on the assumption that the global warehouses uses the first-come first-serve allocation rule. Preferably, these basestock levels were set such that the SLAs are met at the end of each contract period with length \(T\). However, the tactical level planning is not in the scope of this thesis, and therefore it is assumed that the basestock levels are given as input. These fixed basestock levels will be used for the operational level planning for finite horizon SLAs. Then, the number of stockouts that are allowed while still meeting the SLA is set equal to the expected number of stockouts for the given basestock levels. Therefore, \(\tau_j(T)\) is determined by the expected number of stockouts in \((0, T)\). To avoid complex computations while still achieving...
5.2 Allowed number of stockouts

Algorithm 3 NORA Allocation Algorithm

1: for all $i \in I : OH_{i,0} > 1$ do
2: \hspace{0.5cm} while $OH_{i,0} > 1$ do
3: \hspace{1cm} for all $r \in R$ do
4: \hspace{1.5cm} Calculate $NAV^r_i$ as defined in Equation (5.2)
5: \hspace{1cm} end for
6: \hspace{1cm} $r^* \leftarrow \arg \max_r NAV^r_i$
7: \hspace{1cm} for all $j \in J_l$ do
8: \hspace{1.5cm} if $r_j = r^*$ then
9: \hspace{1.75cm} Calculate $NAV_{i,j}$ as defined in Equation (5.1).
10: \hspace{1.5cm} else
11: \hspace{1.75cm} $NAV_{i,j} \leftarrow 0$
12: \hspace{1.5cm} end if
13: \hspace{1.5cm} $j^* \leftarrow \arg \max_j NAV_{i,j}$
14: \hspace{1.5cm} end for
15: \hspace{1cm} $OH_{i,j^*} \leftarrow OH_{i,j^*} + 1$
16: \hspace{1cm} $OH_{i,0} \leftarrow OH_{i,0} - 1$
17: \hspace{1cm} end while
18: end for

A realistic approximation of the actual stockouts, the expected number of stockouts is determined as if each local warehouse operates in isolation. The first step towards the expected number of stockouts is the use of the fillrate per SKU, $\beta_{i,j}(S_i)$, which is the fraction of demand for SKU $i$ at local warehouse $j$ that is fulfilled immediately upon request, i.e. from stock in the local warehouse itself. Similar to Van Aspert (2014, Ch. 4), this is defined by

$$\beta_{i,j}(S_i) := \sum_{x=0}^{S_{i,j}-1} \frac{\lambda_{i,j} \cdot t_i^x}{x!} e^{-\lambda_{i,j} t_i^x}.$$ 

Based on the fillrate for each SKU, the fillrate for the entire local warehouse can be calculated. This is the fraction of the total demand at local warehouse $j$ that is fulfilled immediately upon request. This fillrate is the weighted sum of the fillrates per SKU and is defined by

$$\beta_j(S) := \sum_{i \in I} \frac{\lambda_{i,j}}{\sum_{i \in I} \lambda_{i,j}} \beta_{i,j}(S_i).$$

Then, the fraction of the demand that is not fulfilled immediately is $1 - \beta_j(S)$. Multiplying this fraction with the expected demand over the entire period of length $T$ results in the expected number of stockouts. By noting that this calculation may result in a non-integer, while the number of allowed stockouts must be an integer, $\tau_j(T)$ is rounded down to the nearest integer. Hence, the allowed number of stockouts is defined by

$$\tau_j(T) := \left\lfloor (1 - \beta_j(S)) \cdot \sum_{i \in I} \lambda_{i,j} T \right\rfloor,$$ (5.3)
where we use that the expected demand of SKU $i$ at local warehouse $j$ in $(0,T)$ is $\lambda_{i,j}T$. Note that this performance measure is independent of the total number of failures that occur during the contract.

If the service level agreement would be defined in terms of fillrate rather than an allowed number of stockouts, the performance measure becomes dependent of the total number of failures that occur during the contract. Then, the target would still be a number of allowed stockouts, but this would be determined given the failures that have already occurred and based on the failures that are still expected. This means that one must keep track of multiple variables and update them every time step. Furthermore, the number of failures during relatively small contract periods is often very low, such that only one failure less (or more), puts strictly tighter (or looser) bounds on the stockouts allowed. For example, with a fillrate of 95%, three stockouts are allowed when there are (expected to be) between 41 and 60 failures. However, only 40 failures would allow us to have only two stockouts, which will have a significant impact on the transshipment decisions. Therefore, using the fillrate as an SLA increases the complexity of analyzing the probability of violating a contract significantly.

### 5.3 Steps and Procedures

This section will describe the set-up and characteristics of the simulation. In the development of the heuristics PASA and PASA-P, all local warehouses were analyzed individually, meaning reactive lateral transshipments were not modeled. In real life, the reactive lateral transshipments and emergency shipments do take place, and for that reason they are included in the simulation. Furthermore, the costs for both transshipments are also included. As discussed in Section 3.1, Discrete Event Simulation is used.

In practice, the transshipment times are often relatively short compared to contract periods of several months or even years. Therefore, the transshipment times are modeled to be less than the length of one time step. This way, all transshipments decisions that are made with $t$ time steps left, will take effect in the next period where there are $t-1$ time steps left. This was indicated in Section 4.4, where the transshipment decisions with $t$ time left were made to minimize the costs for a remaining contract period of $t-1$ time steps. In this computational study, one time step is equal to one day.

At the beginning of each day in the simulation, several steps are taken with regard to the order of events occurring. First, it is checked if any new stock has arrived in the global warehouse during the previous day by using Equation (3.2). This incoming supply is first used to fulfill backorders at the global warehouse. Any remaining stock will be allocated to local warehouses. Second, it is checked if failures have occurred in the day before, where Equation (3.1) is used. If this is the case, we check whether it can be fulfilled by the assigned local warehouse, by a reactive lateral transshipment, or...
by an emergency shipment otherwise. Based on the inventory levels resulting from the previous steps, any allocation, and – if included in the heuristic – proactive lateral decisions are made. The new state, resulting from these decisions, become available to the decision maker at the beginning of the next day. All these steps are depicted in Figure 5.1. The final step is only performed if there are more than one day left. Otherwise, decisions would take effect on the moment the contract is finished, so that they can have no influence on the performance in the last time step.

In the simulation, all four heuristics presented in Section 5.1 are modeled simultaneously. This means that failures occur for all four models at the same time, to ensure that these heuristics will be tested under the same circumstances. In the initializing phase, the failure rate $\lambda$ and replenishment rates $\mu$ are generated and based on these values the basestock levels are created. These lead to the allowed number of stockouts as defined in Equation (5.3). At ASML, these basestock levels are determined based on the Algorithm 3 of Kranenburg and Van Houtum (2009). However, the algorithm that they implemented includes more input than the input in the heuristic presented in Chapter 3 and could therefore not be used to generate the basestock levels as input for the simulation. In order to obtain reasonable basestock levels, the failure rates were divided into three groups of equal size, and basestock levels 0, 1 and 2 were assigned. High, medium and low rates correspond to a basestock level of 0, 1 and 2. After all, the number of allowed stockouts is
based on these basestock levels and not the other way around. If, in practice, the correct basestock levels can be determined based on actual contract requirements, these can be translated into the actual number of stockouts that is allowed to meet the contract.

The last part of the initialization is to determine the starting inventory levels. For each SKU $i \in I$ at each local warehouse $j \in J_l$, the on-hand inventory level at the beginning of the contract period is determined by picking a random number between 0 and $S_{i,j}$, based on the steady-state probabilities for that $M/M/S_{i,j}/S_{i,j}$ system for a realistic starting inventory. Taking the initial on-hand inventory levels different from the basestock levels is common in spare parts systems, as in practice this is often the case. In this situation, the global warehouse should know where to send new parts for the FCFS model. A sequence is generated for each SKU by randomly generating a list where each local warehouse is included $S_{i,j} - OH_{i,j}$ times for SKU $i \in I$ and local warehouse $j \in J_l$.

5.3.1 Verification

The simulation model was verified in a number of ways. First of all, whenever it was possible, results from the simulation were compared with theoretical results. For instance, Ferrante (2009) provided an expression for the expected number of losses in a $M/M/1/1$ system, which is an upper bound to the probability of one loss in a $M/M/1/1$ as possible output in my simulation. This could be checked without difficult computations. Other analytic results, for example the sum of zero, one and two losses should be less than one, could be also be verified. For many functions in the simulations, many tests could be generated by varying input parameters over their acceptable range to check if the model provided the correct output. This was done for random instances, but also for extreme instances, in which there was only one option. Furthermore, many calculations were checked manually and it was checked whether the output was realistic by using common sense.
In this chapter, we present a computational study to evaluate the performance of the four heuristics discussed in the previous chapter. First, the test beds and objectives of the computational study are introduced. Subsequently, the numerical results are presented and discussed.

6.1 Test beds and objectives

We are interested in the potential performance improvements for networks similar to the ASML network as a result of the developed heuristics PASA and PASA-P. Therefore, this computational study serves three objectives. First, we want to determine how the performance of the allocation decisions of NORA and PASA compares against the FCFS heuristic. Second, we want to quantify the added value of performing proactive lateral transshipments on top of the allocation decisions by PASA, i.e., comparing the performance of PASA and PASA-P. These two objectives are based on different parameter settings. Therefore, the third objective is to investigate how the answers to the previous two objectives are influenced by different parameter settings.

We note that the time complexity of the PASA-P heuristics increases fast for an increasing number of SKUs, an increasing number of local warehouses, or both. Furthermore, in a spare parts network, failure rates are commonly quite low, such that many simulation runs are needed to find significantly small confidence intervals for the total cost. Moreover, the objective is to find the potential of using the PASA and PASA-P model in practice. This potential can already be tested for small instances with multiple regions with multiple local warehouses and multiple SKUs. For these reasons, we resort to small test instances and a limited number of testbeds.

We choose a small instance, that is large enough to show the potential improvements of the heuristics, that consists of four local warehouses spread over two regions. Each region consists of two local warehouses such that the effect of reactive lateral transshipments is included. For these reactive lateral transshipments, the assumption is made that all local warehouses are mains, such that at least one other local warehouse in the region can be checked for supplying via a reactive lateral transshipment before immediately resorting to the global warehouse for an emergency shipment.
Furthermore, asymmetric failure rates are modeled as this is often the case in practice. The penalty costs are assigned based on total failure rates: the highest penalty costs will be assigned to the local warehouse with the highest total failure rate, the second highest penalty costs to the local warehouse with the second highest total failure rate and so on.

Initially, a small computational study was performed to give a first indication of the effect of the parameter settings. This way, some parameters are fixed to one value, which enables us to reduce the number of test instances. From this small study, it could already be seen that the cost for reactive lateral transshipments and the cost for emergency shipments have little effect on the performance of the heuristics. Additionally, the replenishment rate is kept constant, while the failure rates are varied to test varying values of the load for repair and production. For these parameters, the value is fixed as shown in Table 6.1. We wish to emphasize that all numbers, e.g. for costs, failure rates and supply lead times, are randomly chosen and fictive, but believed to be realistic for spare part networks. Furthermore, in such a small network, the iron stock only leads to higher costs. As there are already few parts for each SKU in the entire system, holding parts in the global warehouse leads to a higher number of violated contracts. Hence, the iron stock is omitted in this computational study for a fair comparison between the four heuristics.

Table 6.1: Base parameter values for computational study

<table>
<thead>
<tr>
<th>Input parameter</th>
<th>Base value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Number of warehouses, including the global warehouse, (</td>
<td>J</td>
</tr>
<tr>
<td>2 Number of regions, (</td>
<td>R</td>
</tr>
<tr>
<td>3 Number of SKUs, (</td>
<td>I</td>
</tr>
<tr>
<td>4 Supplier replenishment rate for SKU (i \in I, \mu_i) (equal to 1/t^s_i, (U[0.04, 0.08]) per day)</td>
<td></td>
</tr>
<tr>
<td>5 Cost for reactive lateral transshipments, (c^\text{rlt}_{j,k}) for all (j, k \in J_l) (in euros)</td>
<td>800</td>
</tr>
<tr>
<td>6 Cost for emergency shipments, (c^\text{em}_j) for all (j \in J_l) (in euros)</td>
<td>1500</td>
</tr>
</tbody>
</table>

The remaining parameter settings were chosen such that they represent realistic situations. Overall, failure rates in spare parts networks are low, but this can vary among different industries. Therefore, three values from the uniform distribution are included in the test bed. The failure rates for each SKU \(i \in I\) at each local warehouse \(j \in J_l\) are drawn from the uniform distribution independent of the other failure rates.

Stock allocations and proactive lateral transshipments cost either 250, 500 or 750 euro, ensuring that this amount is always smaller than the cost for a reactive lateral transshipment or an emergency shipment. Furthermore, in different industries, different penalty costs may be applied for not meeting the required contract performance. We test values from three different uniform distributions, where the penalty cost for each local warehouse is drawn from this distribution independently of the others.

Multiple contract durations are included to test the effect of shorter or longer periods. It is well
known that shorter periods provide the benefit that not many failures can occur during a particular review horizon, increasing the chance of meeting the required performance. Long reviews increase the probability of many failures, but have the benefit that they have a longer time to recover if the system is not performing well at a certain point in time (Thomas, 2005). This phenomenon is investigated in this computational study. An overview of all parameter settings is provided in Table 6.2.

Table 6.2: Parameter settings the computational study

<table>
<thead>
<tr>
<th>Input parameter</th>
<th>No. of choices</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure rate, ( \lambda_{i,j}, \forall i \in I, j \in J_t ) (per day)</td>
<td>3</td>
<td>( \alpha \cdot U[0.001, 0.1] ) for ( \alpha = {0.5, 1, 1.5} )</td>
</tr>
<tr>
<td>Cost for allocation and proactive lateral transshipment, ( c^{tr} ) (in euros)</td>
<td>3</td>
<td>250, 500, 750</td>
</tr>
<tr>
<td>Penalty cost for contract violation, ( c^{pen}_{j}, \forall j \in J_t ) (in euros)</td>
<td>3</td>
<td>( \beta \cdot U[60K, 120K] ) for ( \beta = {0.5, 1, 1.5} )</td>
</tr>
<tr>
<td>Length of the contract period, ( T ) (in days)</td>
<td>3</td>
<td>30, 60, 90</td>
</tr>
</tbody>
</table>

We will perform two simulation studies to test the different objectives mentioned above. Simulation study 1 only includes FCFS, NORA and PASA to investigate the effect of different allocation policies. The computation times of these three heuristics is relatively low compared to PASA-P, such that more instances can be tested. Therefore, all combinations between the parameters given in Table 6.2 are investigated and presented, i.e. 81 instances. In simulation study 2, also PASA-P is included to test the additional effect of proactive lateral transshipments. Taking the (higher) computational time into account, one base case will be considered, this is \( (\alpha, c^{tr}, \beta, T) = (1, 500, 1, 60) \). From here, one parameter will be varied at a time such that we have 9 instance in total for this second simulation study.

To evaluate the effectiveness of the four heuristics, the relative difference between the total cost obtained by the different heuristics is measured, and is given by

\[
\%VAL^x(n) = 100 \cdot \frac{g^n - g^x}{g^x},
\]

where \( g^x \) denotes the total cost for the heuristic \( x \) that is used as benchmark, and \( g^n \) denotes the total cost for heuristic \( n \). These costs are obtained by performing the Discrete Event Simulation study as sketched in the previous section. First, the performance of the FCFS heuristic is taken as a benchmark – because we expect this heuristic to perform the worst – and is compared to the performance of the other three heuristics. That is, evaluating \( \%VAL^1(n) \) for \( n \in \{2, 3, 4\} \).
6.2 Numerical results

All four heuristic approaches were programmed as multi-threaded applications in MATLAB. The computations were carried out on multiple PCs running Windows (64 bit). They had different processors and memory, among which one PC with Intel Dual Core @ 1.60 GHz and 8GB RAM, one PC with Intel Quad Core @ 2.50 GHz and 8GB RAM and two additional PCs with Intel Dual Core @ 3.20 GHz and 4GB RAM. The reason for using multiple machines is the computation time for all the instances, mainly due to the computational efforts for the local search for PASA-P, and the limited time available for retrieving results.

The performance of the heuristics is evaluated by using 2,000 simulation runs for each instance in the first simulation study and 1,000 simulation runs in the second simulation study. As mentioned in Section 5.3, this number of simulation runs are needed for creating a small confidence interval, while the computational time is still practical. We find that the confidence intervals are not small enough, such that the confidence intervals of the different heuristics are not distinct. The conclusions that are drawn from the results are not shown to be significant. Nonetheless, we can clearly see an improvement of NORA over FCFS, PASA over NORA and PASA-P over PASA for the majority of the instances. This already indicates that the two proposed heuristics have the potential to improve the contract performances, although confidence intervals are not distinct, and the results are not shown to be significant.

6.2.1 Simulation study 1

As mentioned above, the first simulation study investigates the effects of different allocation methods, i.e. FCFS, NORA and PASA. The results for the first simulation study are summarized in Table 6.3. Table C.1 in Appendix C shows the 95% confidence intervals for each instance in simulation study 1.

Before going into detail about the effect of each parameter, we summarize the main observations drawn from this table below:

- Table 6.3 shows that both the NORA and PASA heuristic perform well in comparison to the FCFS method. The difference between the NORA heuristic and the PASA heuristic is not very large in some cases, but we can still conclude that, on average, the PASA heuristic outperforms both the FCFS heuristic and the NORA heuristic.

- Although the smallest reduction in terms of total cost per contract period is only a reduction of 1.90%, we see that for low failure rates the highest cost reduction is achieved. This shows potential for the PASA heuristic when implemented at ASML, as the failure rates here are very low.
Table 6.3: Summary of the computational results for simulation study 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Total costs FCFS (€)</th>
<th>Average % $VAL^1(n)$</th>
<th>2 NORA</th>
<th>3 PASA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure rate coefficient, $\alpha$</td>
<td>0.5</td>
<td>$0.72 \cdot 10^5$</td>
<td>7.52</td>
<td>15.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>$2.31 \cdot 10^5$</td>
<td>5.26</td>
<td>6.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>$4.13 \cdot 10^5$</td>
<td>1.90</td>
<td>1.96</td>
<td></td>
</tr>
<tr>
<td>Cost for allocation &amp; proactive</td>
<td>250</td>
<td>$2.40 \cdot 10^5$</td>
<td>3.56</td>
<td>4.50</td>
<td></td>
</tr>
<tr>
<td>lateral transshipments, $c_{tr}$</td>
<td>500</td>
<td>$2.36 \cdot 10^5$</td>
<td>3.52</td>
<td>4.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>$2.40 \cdot 10^5$</td>
<td>3.56</td>
<td>4.82</td>
<td></td>
</tr>
<tr>
<td>Penalty cost coefficient, $\beta$</td>
<td>0.5</td>
<td>$1.51 \cdot 10^5$</td>
<td>3.39</td>
<td>3.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>$2.40 \cdot 10^5$</td>
<td>3.18</td>
<td>4.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>$3.25 \cdot 10^5$</td>
<td>3.89</td>
<td>5.08</td>
<td></td>
</tr>
<tr>
<td>Length of the contract period, $T$</td>
<td>30</td>
<td>$1.90 \cdot 10^5$</td>
<td>3.42</td>
<td>6.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>$2.37 \cdot 10^5$</td>
<td>3.73</td>
<td>4.57</td>
<td></td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>$2.88 \cdot 10^5$</td>
<td>3.48</td>
<td>3.48</td>
<td></td>
</tr>
</tbody>
</table>

- Both for higher failure rates and longer contract periods, the difference in performance between NORA and PASA becomes very small. This might be due to the fact that for an increase for these parameters, the allowed number of stockouts increases as well. However, the maximum number of stockouts that can be achieved during a contract period based on the calculations of PASA is only $2 \cdot |I| = 10$ for this particular network. Then, if the remaining number of allowed stockouts is larger than 10, the probability of violating the contract is zero for all local warehouses and new parts are assigned to the lowest numbered local warehouse where the inventory level is lower than the basestock level. This results in many parts being send to the same local warehouse(s), and does not benefit the performance of PASA. In such cases, allocation based on the NORA or random allocation might mitigate these effects.

The contribution of all cost factors to the total cost per contract period were also investigated. The average relative distributions of the cost due to allocation, reactive lateral transshipments, emergency shipments and contract violations are summarized in Table 6.4.

Table 6.4: Contribution to total cost per contract period for simulation study 1 (in %)

<table>
<thead>
<tr>
<th>Costs due to:</th>
<th>Allocation</th>
<th>Reactive lateral transshipments</th>
<th>Emergency shipments</th>
<th>Violation penalties</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 FCFS</td>
<td>4.8</td>
<td>2.0</td>
<td>19.6</td>
<td>73.5</td>
</tr>
<tr>
<td>2 NORA</td>
<td>5.2</td>
<td>2.1</td>
<td>19.7</td>
<td>73.0</td>
</tr>
<tr>
<td>3 PASA</td>
<td>5.0</td>
<td>2.2</td>
<td>20.7</td>
<td>72.1</td>
</tr>
</tbody>
</table>
We can observe that the differences are not very large when comparing the three heuristics. The biggest difference can be seen for the costs due to violating contracts, where PASA is able to reduce these cost by 1.4% on average compared to FCFS. Note that the costs due to emergency shipments for PASA are higher than for FCFS and NORA. This might be due to the fact that FCFS, unintentionally, and NORA, intentionally, divide the parts over the different regions more evenly than PASA might, which might reduce the number of emergency shipments. In absolute values, the average total cost is $2.32 \cdot 10^5$ euro (average over FCFS, NORA and PASA as the cost distribution is rather similar). More specifically, the costs due to allocation, reactive lateral transshipments, emergency shipments and contract violations are on average $1.16 \cdot 10^4$, $4.88 \cdot 10^3$, $4.64 \cdot 10^4$ and $1.69 \cdot 10^5$ euros, respectively.

It is also interesting to investigate the effect of varying the parameters in this test bed. For changes in $\alpha, \epsilon^r, \beta$ and $T$, the results are shown in Figure 6.1.

The top-left figure in Figure 6.1 shows that the total costs for all three heuristics increase for increasing failure rates and the average costs for all three heuristics increase at approximately the same rate. This is not remarkable since more failures lead to more available stock in the global
6.2 Numerical results

warehouses. These parts need to be allocated to the local warehouses with cost $c^{tr}$ for each shipment. Furthermore, for increasing failure rate the improvement of PASA, relative to FCFS and NORA, decreases. This is most likely due to the observation already made about the increasing number of allowed stockouts.

For all three heuristics the total costs decrease for increasing the transshipment cost from 250 to 500 euros. Increasing the transshipment cost even further, from 500 to 750 euros, results in an increase of the average total cost. The improvement of PASA and NORA over FCFS remains constant for different transshipment costs. This is no surprise as these three heuristics allocate the parts following their own logic, independent of the transshipment costs. Nonetheless, one would expect increasing total costs for increasing transshipment costs as an equal number of parts is being allocated, but for higher costs. Based on the top-right figure, no clear explanation can be found for these relatively constant average costs. A more detailed analysis is necessary to determine the origin of this effect.

The average total costs per period increase linearly with the penalty cost for not meeting a contract for all heuristics. Increasing or decreasing the penalty costs does not change anything about the FCFS and NORA heuristics, which means that on average the number of local warehouses for which the number of allowed stockouts is exceeded will remain constant for increasing penalty costs. Then, the total cost caused by penalty costs will increase at the same rate. As NORA is somewhat better at avoiding contract violation than FCFS as seen in Table 6.4, the cost will grow at a slightly lower rate as well. PASA is able to avoid even more contract violations, resulting in even lower costs.

Naturally, it holds for all four heuristics that an increased duration of the contract period results in increased average costs per period, as also shown in the bottom-right figure. More time steps in each period lead to more failures, which consequently lead to more stock allocation for replenishments from the supplier. Also, more reactive lateral transshipments and emergency shipments are performed. Furthermore, for increasing contract duration the improvement of PASA over FCFS and NORA decreases, which is again most likely due to the increasing number of allowed stockouts as observed before.

A more detailed look at the number of contracts that is violated is shown in Table 6.5. The results depicted in Table 6.5 show that the NORA heuristics reduces the number of contracts that are violated for all parameter settings. Approximately 0.08 contract violations are avoided per contract period. PASA reduces the number of violated contract even further by approximately 0.06 contracts per contract period. Note that these numbers remain relatively constant for all parameter changes.

Looking at the absolute values for contract violations, we see that an increased failure rate results in the highest increase in violated contracts. This might be explained by a higher number of backorders at the supplier for a higher failure rate. This means less replenishments to the local
warehouses as the global warehouse must first remove these backorders. This leads to lower number of parts in the local warehouses, which makes a contract more volatile to violation. Although the number of allowed stockouts is adjusted to this higher failure rate, the global warehouse and the supplier still work the same, leading to more backorders.

For longer contract duration, we see that the number of contract violations also increases. One could argue that longer contract durations allow for more time to make up for poor contract performance in the previous time steps. However, once a contract is violated there is no way one can recover from this violation, unlike for example in the fillrate performance measure. In essence, for longer contract duration there are more opportunities to violate the contracts, and because of high variance in the total failures per contract period, this seems to occur indeed.

### 6.2.2 Simulation study 2

Hereafter, the results for simulation study 2 will be presented. In this simulation study, PASA-P will also be included to investigate the effect of proactive lateral transshipments on the performance. All four heuristics are included in this analysis, but the focus will be on the improvement from PASA to PASA-P. Nonetheless, similar tables will be presented as in the comparison between FCFS, NORA and PASA, but these are less accurate as it is the result of only one instance instead of taking the average over 27 instances as was done in simulation study 1. However, the goal of the results shown here is to give a clear indication of the possible improvements of PASA-P over PASA relative to the improvements that were already made by NORA and PASA. Table C.2 in Appendix C shows the 95% confidence intervals for each instance in simulation study 2.
6.2 Numerical results

The results in terms of costs and number of violated contracts for simulation study 2, including the PASA-P heuristic, are summarized in Table 6.6 and Table 6.7, respectively.

Table 6.6: Summary of the computational results for simulation study 2

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Total cost FCFS (€)</th>
<th>Average %VAL*(n) heuristics n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure rate coefficient, α</td>
<td>0.5</td>
<td>0.59 · 10^5</td>
<td>9.26, 18.8, 42.5</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.24 · 10^5</td>
<td>5.31, 5.16, 12.8</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.17 · 10^5</td>
<td>2.21, 1.87, 2.40</td>
</tr>
<tr>
<td>Cost for allocation &amp; proactive lateral transshipments, c_tr</td>
<td>250</td>
<td>2.04 · 10^5</td>
<td>7.42, 8.41, 16.5</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>2.24 · 10^5</td>
<td>5.31, 5.16, 12.8</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>2.50 · 10^5</td>
<td>3.76, 5.25, 11.3</td>
</tr>
<tr>
<td>Penalty cost coefficient, β</td>
<td>0.5</td>
<td>1.24 · 10^5</td>
<td>5.90, 5.82, 14.1</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.24 · 10^5</td>
<td>5.31, 5.16, 12.8</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>2.69 · 10^5</td>
<td>5.92, 8.09, 16.2</td>
</tr>
<tr>
<td>Length of the contract period, T</td>
<td>30</td>
<td>2.03 · 10^5</td>
<td>5.16, 8.75, 19.1</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>2.24 · 10^5</td>
<td>5.31, 5.16, 12.8</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>2.80 · 10^5</td>
<td>5.13, 3.38, 10.1</td>
</tr>
</tbody>
</table>

Table 6.7: Summary of the number of violated contracts for simulation study 2

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>FCFS</th>
<th>NORA</th>
<th>PASA</th>
<th>PASA-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure rate coefficient, α</td>
<td>0.5</td>
<td>0.40</td>
<td>0.35</td>
<td>0.28</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.00</td>
<td>1.91</td>
<td>1.85</td>
<td>1.67</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>3.37</td>
<td>3.29</td>
<td>3.27</td>
<td>3.25</td>
</tr>
<tr>
<td>Cost for allocation &amp; proactive lateral transshipments, c_tr</td>
<td>250</td>
<td>1.80</td>
<td>1.66</td>
<td>1.59</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>2.00</td>
<td>1.91</td>
<td>1.85</td>
<td>1.67</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>2.02</td>
<td>1.94</td>
<td>1.85</td>
<td>1.69</td>
</tr>
<tr>
<td>Penalty cost coefficient, β</td>
<td>0.5</td>
<td>1.63</td>
<td>1.50</td>
<td>1.43</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>2.00</td>
<td>1.91</td>
<td>1.85</td>
<td>1.67</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>1.66</td>
<td>1.56</td>
<td>1.47</td>
<td>1.30</td>
</tr>
<tr>
<td>Length of the contract period, T</td>
<td>30</td>
<td>1.98</td>
<td>1.88</td>
<td>1.79</td>
<td>1.58</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>2.00</td>
<td>1.91</td>
<td>1.85</td>
<td>1.67</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>2.31</td>
<td>2.18</td>
<td>2.15</td>
<td>1.95</td>
</tr>
</tbody>
</table>

The main observations drawn from both tables can be summarized as follows:

- We see some changes between the cost reductions for the NORA and PASA heuristic of Table 6.3 and Table 6.6. In Table 6.6, NORA sometimes yields better results than PASA. However, this might be instance specific as in simulation study 2 these costs are the result of testing only one instance, where in simulation study 1 these costs are the average over all instances for that fixed parameter (1 instance vs. average of 27
instances). However, regarding only the number of contracts that have been violated in these instances, PASA is able to achieve a lower number of violated contracts, even though this does not seem to always result in lower costs. Or stated otherwise, lower costs does not imply a lower number of contract violations.

- The main focus of this second simulation study was to investigate the added value of proactive lateral transshipments. As can be seen, both the NORA and PASA heuristic lead to both cost reductions as well as reductions in the number of contract violations already, but the addition of proactive lateral transshipments achieves even higher reductions. The percentage of costs that is saved relative to the FCFS heuristic, on top of what both NORA and PASA achieved, is significant. For instance, if the current situation at ASML is a situation with \( \alpha = 0.5 \), PASA-P is able to reduce the NORA costs by almost 37% (i.e. \( \%VAL^2(4) \)). Furthermore, PASA-P performs well in terms of the number of contract violations as well.

- We note that the relative performance of PASA-P over FCFS decreases for longer contract durations. This may be due to proactive lateral transshipments in the early phases of the contract period. A part may be transshipped with 88 days left if this reduces the expected costs by only one euro. A threshold for proactive transshipments might be incorporated. For example, this threshold could be a cost reduction of a certain amount, or a cost reduction of a certain amount that is based on the time left in the contract.

Again, we are interested in the division of the total costs over the different cost factors. The average absolute costs are summarized in Table 6.8. It can be observed that the differences in costs for allocation, reactive lateral transshipments and emergency shipments between the four heuristics are relatively small. The main reduction in cost is for reducing penalty costs for contract violations. Here, NORA and PASA are already effective, but adding proactive lateral transshipments shows an even more significant cost reduction for the penalty costs. On average, the NORA, PASA and PASA-P heuristic reduce the costs of FCFS by 4.8%, 5.7% and 12.5%, respectively. The main contributor to these reductions is the reduction of penalty cost, which is good for more than 85% of the cost reduction.

Table 6.8: Average absolute costs due to different cost factors per contract period for simulation study 2 (\( \cdot 10^4 \) €)

<table>
<thead>
<tr>
<th>Costs due to:</th>
<th>Allocation</th>
<th>Reactive lateral transshipments</th>
<th>Emergency shipments</th>
<th>Violation penalties</th>
<th>Total costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 FCFS</td>
<td>1.25</td>
<td>0.54</td>
<td>4.05</td>
<td>16.72</td>
<td>22.55</td>
</tr>
<tr>
<td>2 NORA</td>
<td>1.29</td>
<td>0.52</td>
<td>3.89</td>
<td>15.74</td>
<td>21.45</td>
</tr>
<tr>
<td>3 PASA</td>
<td>1.24</td>
<td>0.55</td>
<td>4.08</td>
<td>15.38</td>
<td>21.26</td>
</tr>
<tr>
<td>4 PASA-P</td>
<td>1.27</td>
<td>0.59</td>
<td>3.96</td>
<td>13.91</td>
<td>19.74</td>
</tr>
</tbody>
</table>
Regarding the relative contribution of the different factors to the total costs, as depicted in Figure 6.2, the same observations can be made. We can clearly see that the contribution of penalty costs to the total cost decreases by using NORA, PASA and PASA-P rather than the FCFS heuristic.

Figure 6.2: Contribution of different cost factors as fraction of the total cost per contract period for simulation study 2
Chapter 7

Conclusions and Discussion

This chapter discusses the main results found in this thesis. The limitations of this research are discussed and recommendations for future research are provided in Section 7.2.

7.1 Main results

In this research, a multi-item, multi-location, two-echelon spare parts model is considered. The model consists of one global warehouse and multiple local warehouses around the world. Both warehouse types use a basestock policy for ordering parts at the supplier and the global warehouse, respectively. This system is a mixture of an adjusted lost sales and a backorder model, including reactive lateral transshipments and emergency shipment. For each local warehouse a service level agreement (SLA) is set up that assures a certain service level within a fixed and finite period of time, called the contract period. The duration of the contract period is considered to be a finite time period, often a few months or years.

The possibilities to reduce the overall costs by increasing the contract performances are investigated. More specific, it is investigated how proactive lateral transshipments and smarter allocation rules can be used to reduce the expected costs until the end of the contract period. This is done by reducing the risk of possible stockouts in order to increase the probability of meeting an SLA, which is defined by the number of stockouts that are allowed during the contract period.

From literature, it can be concluded that the intersection between finite horizon review period and both the proactive lateral transshipments and the stock allocation decisions needs further investigation. Since this spare parts network can easily be translated to a queueing system, where finite time periods can be considered, we resorted to transient loss systems for the development of two new heuristics that deal with this gap. These transient loss systems are used to find the distribution for the number of stockouts, which is used to calculate the probability of exceeding the allowed number of stockouts until the end of the contract period for each warehouse. Then, the expected costs until the end of the contract period is derived, and two optimization problems that minimize these costs are formulated for our two heuristics.
In practice, allocation-only policies are often used in the operational level decisions. Therefore, the first of our two heuristics, called PASA, only makes allocation decisions. Our second heuristic, PASA-P, is an extension of the first by also allowing proactive lateral transshipments. These two heuristics are performance anticipating heuristics that take the current state of the system into account, e.g. the on-hand inventory levels in comparison to the basestock levels, the remaining number of allowed stockouts and the time until the end of the contract period.

The performance of our proposed heuristics is compared with the FCFS heuristic and the NORA heuristic, which is currently used in practice (i.e. by ASML) across a large test bed of industrial size and where decisions are based on the steady-state Erlang blocking probability. The comparison of the performance is done for small networks and with a limited number of instances due to the computational complexity of simulation. We found that the NORA, PASA and PASA-P heuristics showed a significant improvement compared to the FCFS heuristics. The maximum cost reductions in the second simulation study were 9%, 18% and 42% for NORA, PASA and PASA-P, respectively. Although NORA already showed an improvement over FCFS, the potential of the two proposed heuristics should be acknowledged. Especially, the proactive lateral transshipment seem to be of great value in reducing the average total costs, by reducing the violated contracts.

This research is the first to investigate the combination of finite horizon review periods with stock allocation as well as the combination of finite horizon review periods with proactive lateral transshipments. Furthermore, the first two concepts were included in our PASA model, whereas all three were incorporated in the PASA-P heuristic. We show that the our proposed allocation-only heuristic can lead to large cost reductions (up to 18%), while PASA-P shows even more potential for reducing the costs (up to 42%).

Although the main focus of this work is on spare parts networks, it is also relevant for other logistics and distribution networks where similar decisions on operational level are important. An example is the emergency vehicles (ambulance) problem, where it is needless to say that no queues are allowed, due to the nature of the problem. Furthermore, any incoming calls (demand) will be served immediately. If no ambulances are available, they are satisfied by other methods, such as the reactive and emergency shipments in our work. Ferrante (2009) also used these models as the motivation for his work.


7.2 Limitations and recommendations

This section discusses the limitations of this research. Additionally, suggestions for future research are provided.

The first limitation of this thesis is the use of one performance target per local warehouse. This is in contrast with practice, where multiple customers are assigned to one local warehouse and these
customers all have their own service level agreement. Furthermore, setting the target SLA in terms of allowed number of stockouts is not common in spare parts networks. Therefore it is recommended to look for methods that include the contracts of multiple customers per local warehouse in the model. If the SLAs will be set in terms of fill rate, one should keep track of both the total number of failures and the number of losses. Based on the number of failures, the number of allowed stockouts can be determined, or predicted with more certainty than by only using the expected number of losses as input for the model. This would also allow for dynamic changes in the target SLA based on the actual number of failures and losses that have already occurred during the contract period, but it is analytically more complicated.

The second limitation is that known incoming supply is not taken into account. When performing a proactive lateral transshipment, the question arises whether it is smart to perform the transshipment if new supply is arriving within one or two weeks. This information could be taken into account in the decision-making process. However, this information is not available in our model. In steady-state, a solution might be to model the supply lead time as a deterministic variable. The analysis for the number of losses does not change as the Erlang loss system is insensitive to the distribution of the service times beyond its mean (Ross, 1996), in steady-state. We do not know how different service time distributions influence the time-dependent behavior of an Erlang loss systems, and this must be investigated. Another solution enabling us to take into account incoming supply is finding the distribution of losses based on discrete time calculations. This will most likely result in recursive formulas that might be more easy to solve, which furthermore result in reduction of computational time. Although the heuristics proposed here will only use the probability generating functions, which are given for general $M/M/K/K$ systems in Section 4.2.4, discrete time calculations could include incoming supply and might be an interesting direction for further research.

The third limitation is the assumption that the fixed basestock levels are determined on tactical planning level and they are based on steady state performance of the system. Although the use of tactical parameters in operational level planning problems is a common approach in practice, the drawback is that performance then depends on how well the tactical parameters are determined. Sequential and simultaneous approaches resolve this issue by using a simple rule that uses real-time information to control operational interventions and optimizes its parameters at a higher planning level either sequentially or simultaneously with tactical planning decisions. Because joint optimization of these problems is highly complex, most papers that consider a sequential or simultaneous approach study small scale problems or problems with more restrictive assumptions. However, it would still be recommended to investigate the effect of integrated tactical and operational planning.

The fourth limitation is that only a small network was used to test the potential of the PASA and PASA-P policies due to computational complexity. It is shown that the potential of the PASA-P heuristic is significant, if compared to the FCFS heuristics as well as to the NORA and PASA policy, but it should still be tested for large systems as the heuristics might behave differently for larger
systems. Testing these large networks is computationally intensive. Furthermore, the calculation of the probability on contract violation becomes more complex for an increasing number of parts. Also, the local search heuristic that is used for proactive lateral transshipments, as implemented in the simulation, does not scale well to complexity. This is due to the fact that the number of possible ‘neighbors’ increases for increasing number of parts and/or local warehouses. A possible solution to cope with the complexity of these large systems was already mentioned in Section 4.3. Therefore, the use of central limit theorem or methods for a two-moment fit should be investigated before implementing the proposed heuristics in large systems.

Another recommendation is to investigate the effect of extending the minimization problems by incorporating decisions for reactive lateral transshipments as well. Although in most cases static rules are in place for these transshipments, there will probably be some cases in which the difference in distance or cost is very small and another local warehouse could be checked first. Here, it might be interesting to investigate the possibilities and effects of choosing the local warehouse to supply a reactive lateral transshipment based on the expected total cost rather than to follow a predefined sequence.


7.2 Limitations and recommendations


APPENDIX A

EXACT LOSS PROBABILITIES FOR A M/M/1/1 LOSS SYSTEM

This appendix presents exact expressions for \( P(L(t) = n|X(0) = x) \) \((n \in \{0, 1, 2\}, \ x \in \{0, 1\})\) by applying Equation (4.5) to the M/M/1/1 system discussed in Section 4.2.2. The expressions for the probability on zero losses provided by Ferrante (2009) are equal to the expressions given below.

\[
P(L(t) = 0|X(0) = 0) = H_0(0,t) = \left(\frac{1}{2} + \frac{2\lambda + \mu}{2\sqrt{4\lambda \mu + \mu^2}}\right) e^{\frac{1}{2}(2\lambda - \mu + \sqrt{4\lambda \mu + \mu^2})t} + \left(\frac{1}{2} - \frac{2\lambda + \mu}{2\sqrt{4\lambda \mu + \mu^2}}\right) e^{\frac{1}{2}(2\lambda + \mu - \sqrt{4\lambda \mu + \mu^2})t}.
\]

\[
P(L(t) = 0|X(0) = 1) = H_1(0,t) = \left(\frac{1}{2} + \frac{\mu}{2\sqrt{4\lambda \mu + \mu^2}}\right) e^{\frac{1}{2}(2\lambda - \mu + \sqrt{4\lambda \mu + \mu^2})t} + \left(\frac{1}{2} - \frac{\mu}{2\sqrt{4\lambda \mu + \mu^2}}\right) e^{\frac{1}{2}(2\lambda + \mu - \sqrt{4\lambda \mu + \mu^2})t}.
\]

\[
P(L(t) = 1|X(0) = 0) = \left. \frac{d}{dz} H_0(z,t) \right|_{z=0} = \left(\frac{-\lambda^2 \mu}{(4\lambda \mu + \mu^2)^{3/2}} + \frac{\lambda t}{4} \left(1 + \frac{2\lambda + \mu}{\sqrt{4\lambda \mu + \mu^2}}\right) \left(1 - \frac{\mu}{\sqrt{4\lambda \mu + \mu^2}}\right) \right) e^{\frac{1}{2}(2\lambda + \mu - \sqrt{4\lambda \mu + \mu^2})t}
\]

\[
+ \left(\frac{\lambda^2 \mu}{(4\lambda \mu + \mu^2)^{3/2}} + \frac{\lambda t}{4} \left(1 - \frac{2\lambda + \mu}{\sqrt{4\lambda \mu + \mu^2}}\right) \left(1 + \frac{\mu}{\sqrt{4\lambda \mu + \mu^2}}\right) \right) e^{\frac{1}{2}(2\lambda + \mu + \sqrt{4\lambda \mu + \mu^2})t}.
\]

\[
P(L(t) = 1|X(0) = 1) = \left. \frac{d}{dz} H_1(z,t) \right|_{z=0} = \left(\frac{-2\lambda^2 \mu}{(4\lambda \mu + \mu^2)^{3/2}} + \frac{\lambda t}{4} \left(1 - \frac{\mu^2}{4\lambda \mu + \mu^2}\right) \right) e^{\frac{1}{2}(2\lambda + \mu - \sqrt{4\lambda \mu + \mu^2})t}
\]

\[
+ \left(\frac{2\lambda^2 \mu}{(4\lambda \mu + \mu^2)^{3/2}} + \frac{\lambda t}{4} \left(1 - \frac{\mu^2}{4\lambda \mu + \mu^2}\right) \right) e^{\frac{1}{2}(2\lambda + \mu + \sqrt{4\lambda \mu + \mu^2})t}.
\]

57
Appendix A. Exact loss probabilities for a $M/M/1/1$ loss system

\[
P(L(t) = 2|X(0) = 0) = \left. \frac{1}{2} \frac{d^2}{dz^2} H_0(z, t) \right|_{z=0} = \frac{1}{2} \left[ \left( -\lambda^2 \mu \cdot \left( \lambda t - \frac{\lambda \mu t}{\sqrt{4\lambda \mu + \mu^2}} \right) + \frac{1}{2} \cdot \left( \frac{-2\lambda + \mu}{2\sqrt{4\lambda \mu + \mu^2}} \right) \cdot \frac{-2\lambda^3 \mu t}{(4\lambda \mu + \mu^2)^{3/2}} \right) + \frac{1}{8} \left( 1 + \frac{2\lambda + \mu}{\sqrt{4\lambda \mu + \mu^2}} \right) \cdot \left( \lambda t - \frac{\lambda \mu t}{\sqrt{4\lambda \mu + \mu^2}} \right)^2 + \frac{2\lambda^3 - \lambda \mu + \mu^2}{(4\lambda \mu + \mu^2)^{5/2}} \right] e^{-\frac{1}{2} \left( 2\lambda + \mu - \sqrt{4\lambda \mu + \mu^2} \right) t} \right]
\]

\[
P(L(t) = 2|X(0) = 1) = \left. \frac{1}{2} \frac{d^2}{dz^2} H_1(z, t) \right|_{z=0} = \frac{1}{2} \left[ \left( \frac{-2\lambda^2 \mu}{(4\lambda \mu + \mu^2)^{3/2}} \cdot \left( \lambda t - \frac{\lambda \mu t}{\sqrt{4\lambda \mu + \mu^2}} \right) + \frac{1}{2} \cdot \frac{\mu}{2\sqrt{4\lambda \mu + \mu^2}} \cdot \frac{-2\lambda^3 \mu t}{(4\lambda \mu + \mu^2)^{3/2}} \right) + \frac{1}{8} \left( 1 + \frac{\mu}{\sqrt{4\lambda \mu + \mu^2}} \right) \cdot \left( \lambda t - \frac{\lambda \mu t}{\sqrt{4\lambda \mu + \mu^2}} \right)^2 - \frac{6\lambda^3 \mu^2}{(4\lambda \mu + \mu^2)^{5/2}} \right] e^{-\frac{1}{2} \left( 2\lambda + \mu + \sqrt{4\lambda \mu + \mu^2} \right) t} \right]
\]
Appendix B

Complexity of the Probability Generating Function of a $M/M/2/2$ Loss System

We consider $H_x(z,t)$ and $s_1(z), s_2(z)$ and $s_3(z)$ as in Section 4.2.3 and note that $s_3(z)$ is the complex conjugate of $s_2(z)$. However, with both $s_2(z)$ and $s_3(z)$ complex numbers, the goal is to show that $H_x(z,t)$ itself is not complex.

$H_x(z,t)$ is written as

$$H_x(z,t) = \sum_{k=1}^{3} e^{s_k(z)t} \left( s_{\nu_k,3}(z) - s_{\nu_k+1,3}(z) \right) g_{x,k}(z) \left( s_1(z) - s_2(z) \right) \left( s_1(z) - s_3(z) \right) ,$$

where

$$g_{x,k}(z) = \begin{cases} s_k(z)^2 + (3 \lambda + 3 \mu - \lambda z)s_k(z) - 2\lambda^2 z + 3\lambda^2 - \lambda \mu z + 3\lambda \mu + 2\mu^2 & \text{for } x = 0, \\ s_k(z)^2 + (3 \lambda + 3 \mu - \lambda z)s_k(z) - \lambda^2 z + 2\lambda^2 - \lambda \mu z + 3\lambda \mu + 2\mu^2 & \text{for } x = 1, \\ s_k(z)^2 + (2 \lambda + 3 \mu)s_k(z) + \lambda^2 + 2\lambda \mu + 2\mu^2 & \text{for } x = 2, \end{cases}$$

Each $g_{x,k}(z)$ is in the form of $s_k(z)^2 + A_x(z)s_k(z) + B_x(z)$, where

$$A_x(z) = \begin{cases} 3\lambda + 3 \mu - \lambda z & \text{for } x = 0, \\ 3\lambda + 3 \mu - \lambda z & \text{for } x = 1, \\ 2\lambda + 3 \mu & \text{for } x = 2, \end{cases}$$

$$B_x(z) = \begin{cases} -2\lambda^2 z + 3\lambda^2 - \lambda \mu z + 3\lambda \mu + 2\mu^2 & \text{for } x = 0, \\ -\lambda^2 z + 2\lambda^2 - \lambda \mu z + 3\lambda \mu + 2\mu^2 & \text{for } x = 1, \\ \lambda^2 + 2\lambda \mu + 2\mu^2 & \text{for } x = 2. \end{cases}$$

For notational convenience, the dependence on $z$ is dropped in the remainder of this appendix.
Note that
\[ e^{s_2t}(s_3 - s_1) \left( s_2^2 + A_x s_2 + B_x \right) + e^{s_3t}(s_1 - s_2) \left( s_3^2 + A_x s_3 + B_x \right) \]
can be rewritten to
\[ -e^{s_2t}(s_1 - s_3) \left( s_2^2 + A_x s_2 + B_x \right) + e^{s_3t}(s_1 - s_2) \left( s_3^2 + A_x s_3 + B_x \right), \]  
(B.1)
from which it can be seen that this factor is asymmetric in \( s_2 \) and \( s_3 \), and it is expected that this factor does not yield a real function. Hence, \( \sum_{k=1}^{3} e^{s_k t} \left( s_{\nu_k,3} - s_{\nu_k+1,3} \right) g_{x,k} \) will probably be complex functions. Evaluating the terms separately, we find

The denominator, \( (s_1 - s_2)(s_1 - s_3)(s_2 - s_3) = 2i \text{Im}(s_2) \cdot ((s_1 - \text{Re}(s_2))^2 + \text{Im}(s_2)^2) = C_{x,1} \cdot i \) for a constant \( C_{x,1} \in \mathbb{R} \). Hence, \( (s_1(z) - s_2(z))(s_1(z) - s_3(z))(s_2(z) - s_3(z)) \) is purely imaginary.

\( s_1 \in \mathbb{R} \), such that \( e^{s_1 t}(s_2 - s_3) \left( s_1^2 + A_x s_1 + B_x \right) \) is purely imaginary as all terms are in \( \mathbb{R} \) except for \( (s_2 - s_3) \). This term is equal to \( s_2 - s_3 = 2i \text{Im}(s_2) \) for complex conjugates, so \( e^{s_1 t}(s_2 - s_3) \left( s_1^2 + A_x s_1 + B_x \right) = C_{x,2} \cdot i \) for a constant \( C_{x,2} \in \mathbb{R} \).

The final step is further analysis of the factors given in Equation (B.1). Here, using (among others) that \( e^{i\phi} + e^{-i\phi} = 2 \cos(\phi) \), \( e^{i\phi} - e^{-i\phi} = 2i \sin(\phi) \) and the fact that \( s_3 = \overline{s_2} \), it can be concluded that
\[ -e^{s_2t}(s_1 - s_3) \left( s_2^2 + A_x s_2 + B_x \right) + e^{s_3t}(s_1 - s_2) \left( s_3^2 + A_x s_3 + B_x \right) = C_{x,3} \cdot i \] for a constant \( C_{x,3} \in \mathbb{R} \), and thus is also purely imaginary.

Combining this, we have:
\[ H_x(z, t) = \frac{(C_{x,2} + C_{x,3})i}{C_{x,1} \cdot i} = \frac{C_{x,2} + C_{x,3}}{C_{x,1}} \in \mathbb{R}, \]
from which we can conclude that \( H_x(z, t) \) is indeed real.
**Appendix C**

**Confidence intervals for the computational study**

Table C.1: The 95% confidence interval for each instance of simulation study 1

| Instance $^1$ $(\alpha, c^\text{tr}, \beta, T)$ | 95% confidence intervals of the average total cost per period for heuristic $n$ (in thousands) | 1 | 2 | 3 |
|---|---|---|---|---|---|
| | FCFS | NORA | PASA | FCFS | NORA | PASA |
| LB | UB | LB | UB | LB | UB | LB |
| 1 | (0.5, 250, 0.5, 30) | 27.74 | 30.78 | 24.96 | 27.83 | 24.00 | 26.83 |
| 2 | (0.5, 250, 0.5, 60) | 37.87 | 40.84 | 35.03 | 37.95 | 32.55 | 35.27 |
| 3 | (0.5, 250, 0.5, 90) | 48.86 | 51.53 | 45.13 | 47.58 | 45.01 | 47.45 |
| 4 | (0.5, 250, 1.0, 30) | 31.93 | 34.67 | 30.90 | 33.56 | 29.28 | 31.88 |
| 5 | (0.5, 250, 1.0, 60) | 63.35 | 67.01 | 59.02 | 62.61 | 55.43 | 58.71 |
| 6 | (0.5, 250, 1.0, 90) | 64.97 | 68.62 | 60.37 | 63.65 | 58.13 | 61.36 |
| 7 | (0.5, 250, 1.5, 30) | 30.29 | 33.00 | 28.49 | 31.06 | 27.04 | 29.53 |
| 8 | (0.5, 250, 1.5, 60) | 46.24 | 49.11 | 43.49 | 46.26 | 41.63 | 44.31 |
| 9 | (0.5, 250, 1.5, 90) | 73.31 | 76.25 | 68.75 | 71.48 | 66.26 | 69.03 |
| 10 | (0.5, 500, 0.5, 30) | 65.91 | 71.49 | 63.53 | 68.84 | 58.04 | 63.27 |
| 11 | (0.5, 500, 0.5, 60) | 54.15 | 58.88 | 50.37 | 54.78 | 44.06 | 48.34 |
| 12 | (0.5, 500, 0.5, 90) | 66.39 | 72.98 | 63.72 | 68.84 | 58.49 | 63.27 |
| 13 | (0.5, 500, 1.0, 30) | 63.87 | 69.82 | 60.38 | 66.12 | 55.13 | 60.59 |
| 14 | (0.5, 500, 1.0, 60) | 82.36 | 88.40 | 78.37 | 84.19 | 66.45 | 71.92 |
| 15 | (0.5, 500, 1.0, 90) | 94.86 | 100.9 | 87.61 | 93.22 | 78.35 | 83.54 |
| 16 | (0.5, 500, 1.5, 30) | 71.72 | 77.83 | 67.30 | 73.22 | 60.57 | 66.28 |
| 17 | (0.5, 500, 1.5, 60) | 75.44 | 80.98 | 68.86 | 74.20 | 62.25 | 67.18 |
| 18 | (0.5, 500, 1.5, 90) | 89.59 | 94.79 | 82.11 | 86.84 | 75.59 | 80.07 |

$^1$Then, the failures rate distribution is denoted by $\alpha \cdot U[0.001, 0.1]$ and the penalty cost distribution by $\beta \cdot U[20000, 40000]$. 

Continued on next page
Table C.1 – Continued from previous page

<table>
<thead>
<tr>
<th>Instance (α, c, β, T)</th>
<th>95% confidence intervals of the average total cost per period for heuristic n (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>LB</td>
</tr>
<tr>
<td>19 (0.5, 750, 0.5, 30)</td>
<td>105.8</td>
</tr>
<tr>
<td>20 (0.5, 750, 0.5, 60)</td>
<td>56.91</td>
</tr>
<tr>
<td>21 (0.5, 750, 0.5, 90)</td>
<td>102.5</td>
</tr>
<tr>
<td>22 (0.5, 750, 1.0, 30)</td>
<td>79.91</td>
</tr>
<tr>
<td>23 (0.5, 750, 1.0, 60)</td>
<td>63.94</td>
</tr>
<tr>
<td>24 (0.5, 750, 1.0, 90)</td>
<td>76.11</td>
</tr>
<tr>
<td>25 (0.5, 750, 1.5, 30)</td>
<td>96.45</td>
</tr>
<tr>
<td>26 (0.5, 750, 1.5, 60)</td>
<td>67.29</td>
</tr>
<tr>
<td>27 (0.5, 750, 1.5, 90)</td>
<td>119.4</td>
</tr>
<tr>
<td>28 (1.0, 250, 0.5, 30)</td>
<td>123.4</td>
</tr>
<tr>
<td>29 (1.0, 250, 0.5, 60)</td>
<td>116.2</td>
</tr>
<tr>
<td>30 (1.0, 250, 0.5, 90)</td>
<td>141.3</td>
</tr>
<tr>
<td>31 (1.0, 250, 1.0, 30)</td>
<td>72.56</td>
</tr>
<tr>
<td>32 (1.0, 250, 1.0, 60)</td>
<td>154.3</td>
</tr>
<tr>
<td>33 (1.0, 250, 1.0, 90)</td>
<td>214.5</td>
</tr>
<tr>
<td>34 (1.0, 250, 1.5, 30)</td>
<td>102.3</td>
</tr>
<tr>
<td>35 (1.0, 250, 1.5, 60)</td>
<td>154.3</td>
</tr>
<tr>
<td>36 (1.0, 250, 1.5, 90)</td>
<td>171.8</td>
</tr>
<tr>
<td>37 (1.0, 500, 0.5, 30)</td>
<td>140.4</td>
</tr>
<tr>
<td>38 (1.0, 500, 0.5, 60)</td>
<td>223.7</td>
</tr>
<tr>
<td>39 (1.0, 500, 0.5, 90)</td>
<td>230.1</td>
</tr>
<tr>
<td>40 (1.0, 500, 1.0, 30)</td>
<td>140.8</td>
</tr>
<tr>
<td>41 (1.0, 500, 1.0, 60)</td>
<td>250.4</td>
</tr>
<tr>
<td>42 (1.0, 500, 1.0, 90)</td>
<td>328.2</td>
</tr>
<tr>
<td>43 (1.0, 500, 1.5, 30)</td>
<td>145.1</td>
</tr>
<tr>
<td>44 (1.0, 500, 1.5, 60)</td>
<td>249.0</td>
</tr>
<tr>
<td>45 (1.0, 500, 1.5, 90)</td>
<td>269.2</td>
</tr>
<tr>
<td>46 (1.0, 750, 0.5, 30)</td>
<td>278.2</td>
</tr>
<tr>
<td>47 (1.0, 750, 0.5, 60)</td>
<td>331.8</td>
</tr>
<tr>
<td>48 (1.0, 750, 0.5, 90)</td>
<td>350.6</td>
</tr>
<tr>
<td>49 (1.0, 750, 1.0, 30)</td>
<td>255.2</td>
</tr>
<tr>
<td>50 (1.0, 750, 1.0, 60)</td>
<td>253.6</td>
</tr>
<tr>
<td>51 (1.0, 750, 1.0, 90)</td>
<td>335.6</td>
</tr>
</tbody>
</table>

Continued on next page
Table C.1 – Continued from previous page

95% confidence intervals
of the average total cost per period
for heuristic \( n \) (in thousands)

<table>
<thead>
<tr>
<th>Instance ((\alpha, c_tr, \beta, T))</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FCFS</td>
<td>NORA</td>
<td>PASA</td>
</tr>
<tr>
<td></td>
<td>LB</td>
<td>UB</td>
<td>LB</td>
</tr>
<tr>
<td>52 (1.0, 750, 1.5, 30)</td>
<td>300.8</td>
<td>312.7</td>
<td>288.4</td>
</tr>
<tr>
<td>53 (1.0, 750, 1.5, 60)</td>
<td>291.8</td>
<td>305.1</td>
<td>265.5</td>
</tr>
<tr>
<td>54 (1.0, 750, 1.5, 90)</td>
<td>432.7</td>
<td>446.4</td>
<td>407.8</td>
</tr>
<tr>
<td>55 (1.5, 250, 0.5, 30)</td>
<td>194.8</td>
<td>199.0</td>
<td>191.6</td>
</tr>
<tr>
<td>56 (1.5, 250, 0.5, 60)</td>
<td>278.7</td>
<td>282.3</td>
<td>274.8</td>
</tr>
<tr>
<td>57 (1.5, 250, 0.5, 90)</td>
<td>347.4</td>
<td>347.4</td>
<td>344.4</td>
</tr>
<tr>
<td>58 (1.5, 250, 1.0, 30)</td>
<td>200.4</td>
<td>204.8</td>
<td>196.8</td>
</tr>
<tr>
<td>59 (1.5, 250, 1.0, 60)</td>
<td>283.7</td>
<td>289.0</td>
<td>280.9</td>
</tr>
<tr>
<td>60 (1.5, 250, 1.0, 90)</td>
<td>319.2</td>
<td>328.2</td>
<td>310.9</td>
</tr>
<tr>
<td>61 (1.5, 250, 1.5, 30)</td>
<td>179.4</td>
<td>183.9</td>
<td>175.0</td>
</tr>
<tr>
<td>62 (1.5, 250, 1.5, 60)</td>
<td>251.5</td>
<td>257.4</td>
<td>244.9</td>
</tr>
<tr>
<td>63 (1.5, 250, 1.5, 90)</td>
<td>312.9</td>
<td>319.7</td>
<td>303.2</td>
</tr>
<tr>
<td>64 (1.5, 500, 0.5, 30)</td>
<td>426.0</td>
<td>433.7</td>
<td>419.3</td>
</tr>
<tr>
<td>65 (1.5, 500, 0.5, 60)</td>
<td>439.8</td>
<td>450.5</td>
<td>431.0</td>
</tr>
<tr>
<td>66 (1.5, 500, 0.5, 90)</td>
<td>338.5</td>
<td>346.3</td>
<td>331.5</td>
</tr>
<tr>
<td>67 (1.5, 500, 1.0, 30)</td>
<td>427.0</td>
<td>438.4</td>
<td>418.8</td>
</tr>
<tr>
<td>68 (1.5, 500, 1.0, 60)</td>
<td>516.6</td>
<td>527.7</td>
<td>512.8</td>
</tr>
<tr>
<td>69 (1.5, 500, 1.0, 90)</td>
<td>320.8</td>
<td>321.6</td>
<td>294.7</td>
</tr>
<tr>
<td>70 (1.5, 500, 1.5, 30)</td>
<td>387.1</td>
<td>394.8</td>
<td>378.0</td>
</tr>
<tr>
<td>71 (1.5, 500, 1.5, 60)</td>
<td>531.0</td>
<td>538.1</td>
<td>528.2</td>
</tr>
<tr>
<td>72 (1.5, 750, 0.5, 30)</td>
<td>518.6</td>
<td>531.1</td>
<td>511.3</td>
</tr>
<tr>
<td>73 (1.5, 750, 0.5, 60)</td>
<td>617.8</td>
<td>629.4</td>
<td>605.5</td>
</tr>
<tr>
<td>74 (1.5, 750, 0.5, 90)</td>
<td>638.8</td>
<td>653.0</td>
<td>622.4</td>
</tr>
<tr>
<td>75 (1.5, 750, 1.0, 30)</td>
<td>418.1</td>
<td>431.0</td>
<td>407.3</td>
</tr>
<tr>
<td>76 (1.5, 750, 1.0, 60)</td>
<td>517.4</td>
<td>530.6</td>
<td>503.0</td>
</tr>
<tr>
<td>77 (1.5, 750, 1.0, 90)</td>
<td>687.8</td>
<td>695.6</td>
<td>676.6</td>
</tr>
<tr>
<td>78 (1.5, 750, 1.5, 30)</td>
<td>416.3</td>
<td>431.3</td>
<td>404.7</td>
</tr>
<tr>
<td>79 (1.5, 750, 1.5, 60)</td>
<td>539.9</td>
<td>553.7</td>
<td>527.8</td>
</tr>
<tr>
<td>80 (1.5, 750, 1.5, 90)</td>
<td>661.6</td>
<td>672.7</td>
<td>651.0</td>
</tr>
</tbody>
</table>
Table C.2: The 95% confidence interval for each instance of simulation study 2

<table>
<thead>
<tr>
<th>Instance $(\alpha, c^r, \beta, T)$</th>
<th>95% confidence intervals of the average total cost per period for heuristic $n$ (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FCFS</td>
</tr>
<tr>
<td>1 (1.0, 500, 1.0, 60)</td>
<td>217.1 219.6 205.2 220.3 187.9 203.4</td>
</tr>
<tr>
<td>2 (0.5, 500, 1.0, 60)</td>
<td>55.39 57.25 44.79 51.38 31.65 36.44</td>
</tr>
<tr>
<td>3 (1.5, 500, 1.0, 60)</td>
<td>410.3 402.3 399.9 413.5</td>
</tr>
<tr>
<td>4 (1.0, 250, 1.0, 60)</td>
<td>120.4 120.0 113.2 120.3 102.9 110.1</td>
</tr>
<tr>
<td>5 (1.0, 750, 1.0, 60)</td>
<td>259.3 261.9 237.2 256.9 215.1 235.5</td>
</tr>
<tr>
<td>6 (1.0, 500, 0.5, 60)</td>
<td>196.6 196.7 179.4 194.9 162.7 178.7</td>
</tr>
<tr>
<td>7 (1.0, 500, 1.5, 60)</td>
<td>241.3 248.6 227.8 245.3 212.4 230.7</td>
</tr>
<tr>
<td>8 (1.0, 500, 1.0, 30)</td>
<td>196.0 185.6 178.3 191.7 157.2 171.0</td>
</tr>
<tr>
<td>9 (1.0, 500, 1.0, 90)</td>
<td>272.3 258.0 262.5 278.9 243.4 260.5</td>
</tr>
</tbody>
</table>