The impact of spare parts service measures on the performance measures of a wafer fabrication process

Soellaart, D.H.

Award date:
2019

Eindhoven, 18 January 2019

Technische Universiteit Eindhoven University of Technology

Department of Industrial Engineering & Innovation Sciences
Master Operations Management and Logistics

The impact of spare parts service measures on the performance measures of a wafer fabrication process

(Public version)

Written by:
D.H. (Daniel) Soellaart
Student identity number: 0970574

in partial fulfillment of the requirements for the degree of

Master of Science
in Operations Management and Logistics

Supervisors
dr. ir. R.J.I. (Rob) Basten, TU/e, OPAC
dr. ir N.P. (Nico) Dellaert, TU/e, OPAC
ir. D.P.T. (Douniel) Lamghari-Idrissi, ASML

TUE. School of Industrial engineering.
Series Master Theses Operations Management and Logistics

Subject headings*: double buffering, capital goods manufacturing, wafer fabrication process, re-entrant production processes, service level agreements, service contracts, spare part aggregate fill rate, mean waiting time for a spare part.*

Abstract

We are interested in the influence of spare part service measures, i.e. aggregate fill rate and downtime waiting for a spare part (DTWP), on the performance measures of a wafer fabrication process. This process is, among other things, characterized by re-entrant flows that result in more impact of the bottleneck process. This bottleneck process consists of machines that are provided and maintained by ASML, which commissions this research.

The research consists of two parts. First, we simulate a part of the spare part supply chain of ASML in order to show the relationship between the spare part service measures on the time to repair probability distribution. Second, we use this probability distribution to examine its influence on the process performance measures of the wafer fabrication process.

Finally, we conclude that a downtime waiting for a spare part agreement is most beneficial for ASML's customers (the most positive results on the performance measures) and ASML (cost reduction and customer satisfaction). Especially when a more extensive supply chain model is considered, these benefits are more advantageous.

Management summary

In this thesis, we present the results of our study on the influence of spare part service measures (i.e. fill rate and mean waiting time for a spare part) on the performance measures of a wafer fabrication process (i.e. cycle time, work in progress level, and throughput).

ASML, the company this research is executed for, is a provider of lithographic equipment. These machines form the bottleneck within the wafer fabrication process. This process is, among other things, characterized by re-entrant flows, which increase the influence of the bottleneck process. Besides providing these machines, ASML is supporting customers by offering maintenance and spare part services. These spare part services are established in service level agreements (SLAs). In the case that ASML does not meet these agreements, monetary penalties are given. There are multiple service measures, which impact the time to repair probability distribution differently. However, the difference of impact on the production process between these spare part service measures is unknown in the literature.

The customers of ASML want best results on their performance measures. These results are obtained when there is a low variability of work in progress parts in the fabrication process. One of the most important variables that cause this variability is the time to repair probability distribution of the bottleneck machine. Customers of ASML desire to predict the time to repair in order to determine the performance measures and plan their process. Furthermore, a reduction of the time to repair variability leads to a more positive result of the performance measures and thus a higher customer satisfaction for ASML.

So, for both ASML and its customers, it is valuable to know the influence of different spare part service measures on the time to repair probability distribution and thus on the performance measures of a wafer fabrication process. In order to solve this problem, we define the following main research question:

"How do different spare part service measures and their parameters influence the performance measures (cycle time, work in progress level and throughput) of a wafer fabrication process?"

To answer this research question we first investigate the relationship between the different service measures and the time to repair probability distribution, in Part I. Then, we investigate the impact of the different time to repair probability distributions on the performance measures of the wafer fabrication process in Part II.

Part I Service measures

The first part of the research provides an answer to the following research question:

"How do different spare part service measures and their parameters influence the total inventory costs and the probability distribution of the time to repair?"

In order to compare different service measures, we model a part of the spare part supply chain of ASML. First, we determine the base stock levels for each spare part service measure. Next, we develop a simulation model to analyse the following components for all different scenarios: (i) the waiting time on a spare part probability distribution, (ii) the costs and (iii) different variables (e.g. lead times, number of machines, and the spare part supply chain) that might influence the outcome. Finally, based on the time waiting on a part, we calculate the coefficients of variation of the time to repair for every service measure.

We conclude that the service measure *mean time waiting for a spare part* results in the lowest coefficient of variation. Moreover, we conclude that a single stock location results in higher costs and a higher

coefficient of variation of the time to repair in comparison with multi-location spare part networks. This is due to the fact that multiple locations share the risk of spare parts unavailability's and therefore require lower base stock levels. This holds for every situation since a fill rate agreement is not time restricted.

Part II Performance measures
The second part of the research provides an answer to the following research question:

"How do different time to repair probability distributions influence the performance measures of a wafer fabrication process?"

In order to answer this research question, we first describe the wafer fabrication model mathematically. Then, we analyse analytically the repair process of the bottleneck machine. However, this is to complicated to analyse for the whole wafer fabrication process. Therefore, we use a simulation model with the time to repair probability distributions as provided in Part I as input.

Based on Part II, we prove that the statement of Hopp & Spearman (2001), which conclude that a lower coefficient of variation of the time to repair has a positive impact on the performance measures. Since the wafer fabrication process consists of re-entrant loops, the time to repair within the bottleneck process result in more impact on the performance measures. We note that scenario 2a, which has the lowest coefficient of variation, has the most positive impact on the performance measures.

Conclusion and recommendations
As the main conclusion, we find that a new customer should choose the service measure waiting time for a spare part instead of the fill rate service measure, based on the model we researched. The down time waiting part measure results in the lowest coefficient of variation in the time to repair and, therefore, the most positive results on the process performance measures. Furthermore, from the perspective of ASML, a mean waiting for a spare part service agreement results in less required stock and thus lower costs.

Furthermore, for ASML it is important to take into consideration that single stock locations result in higher costs in comparison with locations with another warehouse nearby. If there is a supporting local warehouse, it is best to provide a down time waiting part service measure since this leads to the most positive results on the performance measures of the wafer fabrication process. ASML can include the extra costs for a single stock point into the price for a service level agreement with the customer.

Now, ASML is always sending parts as fast as possible. However, fill rate agreements are not time restricted. Therefore, ASML can consider the emergency and lateral transshipments to plan groups with this agreement. At the moment, ASML is overperforming which leads to unnecessary costs and effort. In the future, better price agreements can be obtained for ASML since the actual performance and impact on the customer is known. Also, ASML is better able to inform the customer, which leads to a higher customer satisfaction.

By applying the mean waiting for a spare part service measure (also known as DTWP), a win-win scenario is achieved for ASML and its customers. For the customer, this leads to the most positive results on the process performance measures, on the one hand. On the other hand, this results in the lower costs and the highest customer satisfaction for ASML.

Preface

Dear reader,

In the spring of 2017, I got in contact with Douniel Lamghari-Idrissi about doing an internship at ASML. The subject we discussed and the enthusiasm of Mr. Lamghari-Idrissi directly appealed to me. This subject was defined as investigating the impact of different spare part service measures on the probability distribution of the time to repair. This immediately caught my attention because of the connection with spare parts supply chain management and operations management within a manufacturing process. Although, I went to Hong Kong in order to do my exchange semester and travel in Asia first. When I came back in March 2018, my subject and supervisor changed. So, I started on this new subject until I found out that another department already researched my assigned topic. After some meetings, I started again at ASML on the original topic with Mr. Lamghari-Idrissi as my ASML supervisor. I would hereby like to express my gratitude to Mr. Lamghari-Idrissi for all the help, interesting insights and opportunities you gave me.

At building 8, I was welcomed by the Service management team, which quickly felt like home. Especially because the team always freed up their schedule to answer questions and help me further. The atmosphere in the team was great, above all during the many football games we played, which I really enjoyed.

In particular, I would like to thank Rob Basten for all the support and the critical feedback you gave me. More than once you helped me to see things differently and helped me to overcome difficult problems. Your eye for detail and clear feedback helped me to improve my report and further develop my academic writing skills.

Furthermore, I would like to thank my friends and especially Jos for his help during my study in Eindhoven. Not only for the evenings that you helped me with my code but also for the conversations about not-work-related aspects. Moreover, I would thank the rest of my family and friends whose attention and support was very important during my life as a student.

Many special thanks go to my parents and sisters (Laura and Mariëtte) for their encouragement, the fun, and advice. I really appreciate that you were not only there to celebrate high peaks but also provide unconditional support when things went less well. Besides, I would like to thank Jan-Willem for all the pleasure and good conversations we had during our road trips to and from 's Hertogenbosch. Finally, I would like to thank my girlfriend Bente for here love, enthusiasm, and support from Portugal.

For now, I hope that this thesis interests you and that you can recognize the effort, joy and personal development I have experienced during my master thesis project.

Best,

Daniel Soellaart

Eindhoven, December 2018

List of Content

List of figures....vii
List of tables....viii
List of variables....ix
1 Introduction....1
2 Problem context....2
2.1 Company description....2
2.2 Semiconductor manufacturers....2
2.2.1 Semiconductor fabrication process....3
2.2.2 Performance measures....3
2.2.3 System availability and time to repair probability distribution....4
2.3 Spare part service management....5
2.3.1 Demand structure of spare parts....5
2.3.2 Supply structure of spare parts....5
2.3.3 Service contracts....6
3 Research design....8
3.1 Problem description....8
3.2 Scientific relevance....9
3.3 Research questions and methodology....11
3.4 The scope of the research....14
4 The influence of spare part service measures on the probability distribution of the time to repair....16
4.1 Spare parts mathematical model description....16
4.1.1 The supply network....17
4.1.2 The costs....18
4.1.3 Spare part service measures....20
4.1.4 The inventory levels....22
4.1.5 Summary of the assumptions about the service measures....22
4.2 The supply lead times in practice....22
4.3 Spare parts service measures case study....24
4.3.1 Input variables of the simulation model....24
4.3.2 Running and validation of the simulation model....25
4.3.3 General findings of the simulation model....27
4.3.4 Main findings of the simulation model....29
4.4 Conclusions and recommendations of part I....32
5 The influence of the time to repair probability distribution on the wafer fabrication process....34

5.1 General information about the wafer fabrication process 34
5.2 Wafer fabrication model description 35
5.2.1 Wafer fabrication process 35
5.2.2 Process steps 36
5.2.3 Summary of the assumptions about the wafer fabrication model 38
5.3 Wafer fabrication case study 39
5.3.1 Analytical modeling of the wafer fabrication process 39
5.3.2 The simulation model and the input variables of the wafer fabrication process 40
5.3.3 Running and validation of the wafer fabrication simulation model 42
5.3.4 Findings of the wafer fabrication simulation model 43
5.4 Conclusions and recommendations of part II 44
6 Conclusions and recommendations 45
6.1 Conclusions 45
6.2 Main limitations of the research 46
6.3 Recommendations and further research 46
6.3.1 Corporate recommendations 46
6.3.2 Academic recommendations 48
6.4 Academic contribution 48
References 49
Appendix A Abbreviations 52
Appendix B Employee references 53
Appendix C Optimization problems 54
Appendix D Fitting lead time probability distributions 55
Appendix E Warm-up times 56
Appendix F Calculation of the number of replications 57
Appendix G Validation 58
Appendix H Input extreme values simulation models 59
Appendix I Comparison of the number of machines at the customer 60
Appendix J Waiting time probability distributions 61
Appendix K Transition matrix Repair process 63
Appendix L Probabilities to be in which phase of the repair process 65
Appendix M Determine the input of the system 66

List of figures

Figure 1 Structure of the research 1
Figure 2 ASML service supply network (Van Aspert, 2014) 6
Figure 3 mean waiting time for a spare part (Van Aspert, 2014) 6
Figure 4 Relationship between spare part service measures and wafer fabrication process 11
Figure 5 Investigated supply models 17
Figure 6 The different scenarios 21
Figure 7 Probability plot: lognormal, exponential and gamma distribution fit on emergency lead time 23
Figure 8 Histogram of emergency time distribution 23
Figure 9 The investigated scenarios 29
Figure 10 Process steps of a layer (Gkorou, et al., 2017) 35
Figure 11 Investigated wafer fabrication process 36
Figure 12 Delay process 37
Figure 13 Flow diagram for repair process model 40
Figure 14 Researched scenarios 45
Figure 15 Histogram plotting local lead time data 55
Figure 16 Histogram plotting lateral time data 55
Figure 17 Histogram plotting emergency time data 55
Figure 18 Warm up time simulation Part I 56
Figure 19 Warm up time simulation Part II 56
Figure 20 WIP level in the process 59
Figure 21 Comparison number of machines 60
Figure 22 Waiting time distribution scenario 1a 61
Figure 23 Waiting time distribution scenario 1b 61
Figure 24 Waiting time distribution scenario 2a 62
Figure 25 Waiting time distribution scenario 2b 62
Figure 26 Waiting time distribution scenario 3a 63
Figure 27 Waiting time distribution scenario 3b 63
Figure 28 Solvers excel 66

List of tables

Table 1 Variables and description ix
Table 2 Fitting probability distributions 23
Table 3 Input data 24
Table 4 Cost input variables 25
Table 5 Influence emergency time 27
Table 6 Influence lateral lead time 28
Table 7 Influence local lead time 28
Table 8 Corresponding fill rate levels to 1% mean waiting time for a spare part level 29
Table 9 Costs per scenario normalized 30
Table 10 Waiting on spare part time distribution parameters 31
Table 11 Time to repair coefficient of variation 31
Table 12 Scenario conclusions 32
Table 13 Repair process 40
Table 14 Input variables type wafers 41
Table 15 Delay mean processing times processes from Akcalt et al. (2001) 41
Table 16 Number of batches per hour that arrive at the system per scenario 42
Table 17 Main results simulation Part II 43
Table 18 Results of the simulation model 44
Table 19 Main conclusion 45
Table 20 Abbreviations 52
Table 21 Employees 53
Table 22 Running simulation Part I 57
Table 23 Running simulation Part II 57
Table 24 Validation simulation model exponential distributed lead times 58
Table 25 Validation simulation model practical lead times 58
Table 26 Output simulation part I by extreme values 59
Table 27 Extreme values output 59
Table 28 Probability to be in a phase per scenario 65

List of variables

Table 1 Variables and description

Variable	Description
σ	Standard deviation
μ	Average
$\theta_{i,n}(S_i)$	Fraction of demand that is not fulfilled by any local warehouse
$A(S)$	Total average availability
$A_i(s_i)$	Average availability of SKU $i \in I$
B	Total number of wafer types
b	Wafer type $b \in B$
$\beta(S)$	Aggregate fill rate
$\beta_i(s_i)$	Item fill rate of SKU $i \in I$
$C(s)$	Total average costs
$C_i(S_i)$	Average costs per time unit for SKU $i \in I$
C_e	Effective process time
c_0	Coefficient of variation in the natural process
c_i^{v}	Value of SKU i of one unit of inventory
c_i^{em}	Emergency transportation costs for SKU $i \in I$
c_i^{Stock}	Stocking cost of SKU i in terms of euros per unit per year
c^{lat}	Cost of a lateral transshipment
c^{loc}	Cost of a local transshipment
c^{rep}	Cost of a replenishment transshipment
c_r	Coefficient of variation (CV) of repair times
CSD	Customer Service Degree, also called aggregate Fill rate in literature
CV	Coefficient of variation in general
$DTWP$	Down time waiting part
I	Number of total different stock keeping units (SKUs)
i	Particular stock keeping unit (SKU)
J	Total number of warehouses
j	Particular warehouse
k	Datapoint of the simulation model
M	Total demand for all SKUs together
m_i	Demand rate for all machines together or arrival rate of SKU $i \in I$
$MTTR$	Mean time to repair
$MTTF$	Mean time to failure
n	Number of the replications simulation model
Q	Specified replication from the simulation model
R	Holding cost rate
r_b	Processing time of wafer type $b \in B$
S	Base stock level for all SKUs
S_i	Base stock level for SKU $i \in I$
SKU i	A particular stock keeping unit
t_i^{loc}	Average time for a local shipment for SKU $i \in I$
t_i^{lat}	Average time for an lateral shipment for SKU $i \in I$
t_i^{em}	Average time for an emergency shipment for SKU $i \in I$
t_i^{rep}	Average time for a replenishment or routine shipment for SKU $i \in I$
t_j	Service time
$W(S)$	Aggregate mean waiting time
$W_i(s_i)$	Mean waiting time for a spare part demand of SKU $i \in I$
$\bar{Y}_k$	Average of multiple replication $k \in K$
Y_k	Output simulation value of replication $k \in K$
Z	Number of machines

1 Introduction

This report contains our study regarding the impact of spare parts service measures on performance measurements within a wafer fabrication process at a semiconductor factory. The capital goods or production systems that are used in these factories are characterized as high-tech, complex and expensive (Enders, 2004). Therefore, large investments are needed and manufacturers expect to use these systems for a long period of time (Jiang, 2012). Furthermore, customers of these expensive capital goods demand a high availability of the system in order to have the best system performance (ASML, 2018a). Original equipment manufacturers (OEMs) deliver, apart from production systems, customer support so that manufacturers can produce as efficiently as possible. This customer support is established in service level agreements (SLAs), which describe the pre-arranged maximal downtime of a system or the number of spare parts that are supplied from a local warehouse (Ge et al., 2018). Not meeting these agreements results in high penalty costs for the OEM. In order to avoid these penalties and to guarantee high system availability, suppliers stock spare parts. These spare parts are further noted as stock keeping units (SKUs).

We execute this research commissioned by ASML, an OEM of photolithography systems, to obtain insights into the influence of spare part service measures on the performance measures in a semiconductor productions system. This is a so-called double buffering problem since both ASML and the customer are keeping stock in order to reduce the variation within the production process. ASML is doing this by stocking spare parts in order to ensure a certain uptime. To reach the same goal, the customer stocks work in progress components (WIP). In literature, this double buffering concept is only researched by Kiesmüller & Zimmerman (2018). Since this subject is rarely researched, the authors investigate a relatively simple production and spare part inventory control process. We continue this research with a more complex production and inventory model whereby multiple spare part service measures are taken into account. Furthermore, next to the WIP level, we investigate the performance measures cycle time and throughput as well. Another important performance measure in the semiconductor industry is the yield. We do not take this performance measure into account since this is linked to the quality of the machine and the process instead of the spare part service measures.

When the influence of the spare part service measures on spare part base stock levels and performance measures of the wafer fabrication process are known, ASML can provide a better service. The customers of ASML are more capable to predict machine downtimes and better organize their process, in order to get better results on their performance measures. As a result, ASML gains a higher customer satisfaction.

In order to find the relationship between the different spare part service measures and the performance measures, this report has the following outline, which we graphically display in Figure 1. Firstly, we describe the problem context where we clarify the background of the problem from the perspective of both ASML and its customers in Chapter 2. Then, we discuss the research design in order to describe the systematic approach used to find an answer to the problem as provided in the problem statement, in Chapter 3. Next, in Chapter 4, we analyse the influence of spare part service measures on the probability distribution of the time to repair within multiple scenarios. After finding this relationship, we investigate the influence of the probability distribution of the time to repair on the wafer fabrication process, in Chapter 5. Finally, in Chapter 6, we discuss the connection between these chapters, which results in the conclusion and recommendations. All abbreviations are given in Appendix A.

Figure 1 Structure of the research

2 Problem context

In this chapter, we cover the background of the research problem. The problem context starts with a description of the company and continues with an explanation of ASML's customers, respectively in the Sections 2.1 and 2.2. ASML and its customers work closely together because aftersales events (i.e. maintenance) are very important in the semiconductor supply chain (Mönch et al., 2017). In literature, directing aftersales events is called service management (Jiang, 2012), which we discuss in Section 2.3.

2.1 Company description

ASML was founded in 1984 and is specialized in lithography systems for the semiconductor industry. This industry is also known for the computer chip industry. Since these chips are used in a wide array of electronic devices, the industry experienced a huge growth in recent years (Banerjea, 2017). The product of ASML, lithographic systems, is the most important part of the chip production process. A lithographic system is modular and can be upgraded and continuously improved. ASML also designs, develops, integrates, markets and services these advanced systems. In addition, ASML has multiple types of lithography systems including Deep Ultraviolet (DUV) and Extreme Ultraviolet (EUV). In the business environment of ASML, these aftersales events become increasingly important for competitive advantage. Also, the direct revenues of the after-sales services are high (Kranenburg, 2006). Furthermore, most of the customers are major global semiconductor manufacturers.

The lithography producer has over 70 locations in 16 countries and the headquarter is based in Veldhoven, the Netherlands (ASML, 2018a). With a market share of 85% in 2017 (ASML, 2018b), ASML is considered to be the world leader. ASML is characterized by its advanced supply chain management (SCM). In other words, there is a close cooperation with the customers and suppliers. This cooperation is based on the strategy of ASML, which consists of technology leadership, customer focus and operational excellence.

The department of SCM is among other things, responsible for the material availability to ASML's factories and systems of the customers. A break in the material availability may lead to system downtime for the customer that could result in serious consequences, i.e. reduced production output (Ge, et al., 2018). To avoid these consequences, ASML and its customers have performance-based contracts. An important section within the performance-based contracts are the service level agreements (SLA's). Not reaching these service levels results in high penalty costs for ASML. The most common performance agreements are based on the mean waiting for a spare part and the fill rate based on spare parts. To avoid penalties, SCM should closely monitor the performance on the service measures and take interventions if needed. In addition, SCM uses spare part safety stocks at the warehouses to be more flexible and to respond to undesirable situations.

2.2 Semiconductor manufacturers

The customers of ASML are the producers of semiconductors and users of the capital-intensive machines of ASML. To provide an insight into the wafer fabrication process, we discuss the general semiconductor fabrication process in Section 2.2.1. Then, in Section 2.2.2, we explain the performance measures within this process. Finally, we provide the importance of system availability within the wafer fabrication process in Section 2.2.3.

2.2.1 Semiconductor fabrication process

The production of semiconductors is a clear example of a high-volume industry with a continuous production process (Hopp & Spearman, 2001). A facility to produce a wafer costs approximately four billion dollars and it takes four to six weeks to fabricate integrated circuits (semiconductor). In order to produce a semiconductor, Monch et al. (2017) identify four different phases in the production of wafers (i.e. building layers, electrically probed, assembled and tested). We focus on the first phase since, on the one hand, Gupta et al. (2006) state that in the semiconductor manufacturing process, building layers is the most costly and time-consuming. On the other hand, ASML produces only machines for this phase. The process to build the different layers (wafer fabrication process) is described by Gupta et al. (2006). These researchers state that the wafer fabrication process is characterized by:

1. A large number of production steps (approximate between 600-800 different steps)
2. Re-entrance of work in process components (recirculating flows)
3. Random equipment failures
4. Sequence-dependent tool setups
5. The need for batch processing tools
6. Expansion of capacity is expensive.

Furthermore, a finished wafer has approximately 20 – 40 layers and for every layer, six sub-processes are executed.

2.2.2 Performance measures

Hopp & Spearman (2001) state that performance measures are indicators of the effectiveness of a process (including throughput, work in process and cycle time). However, in the wafer fabrication process, there is a fourth performance measure (yield). We describe the measures as:

- Throughput (TH), the rate of produced parts within a time section. Hereby the yield and the throughput rate of the bottleneck are of great importance for the manufacturer (Sched, 2011).
- Work In Process (WIP), the average number of semi-finished parts within the entire production line. The WIP consists of, on one hand, parts that are processed within the production processes and, on the other hand, parts that are stocked between the production processes. The number of parts that are stocked between processes is known as WIP buffer. The objective of WIP buffers is to balance the workflow between the different production processes in the production line. The semiconductor processes are executed in clean rooms of class 100, the number of dust particles is four orders of magnitude lower than the air in a normal room, which means that work in progress buffers are expensive (May & Spanos, 2006).
- Cycle Time (CT), the time it takes to produce a product including queuing time, processing time and transportation time.
- Yield, the ratio of wafers that enters the system and the number of wafers that finish the process. we do not focus on the yield since this is a more quality-based performance measure and is not influenced by the availability of spare parts.

The relationship between the three performance measures is given by the law of Little. According to this law, the amount of WIP in a process is equal to the throughput multiplied by cycle time (Little, 1961). Furthermore, variability has, among other things, a huge impact on these performance measures and is defined as "anything that causes the system to depart from regular, predictable behavior" (Hopp & Spearman, 2001, p. 248). There are multiple sources for this variability, including rework, product differentiation, operator unavailability, and machine failures. Especially the latter is important in this research since it is in the ability of ASML to reduce the impact of a machine failure at the customer.

2.2.3 System availability and time to repair probability distribution

At the moment, there is excess demand in the semiconductor market due to an unprecedented revolution in electronic devices (Bahai, 2017). Prime examples that cause this revolution are smart personal electronics, autonomous systems (e.g. automotive industry) and smart factories. Since all semiconductors can be sold due to excess demand, productivity is extremely important for manufacturers. In order to achieve high productivity, system availability is a key factor (Jiang, 2012). One of the largest users of ASML systems even states that availability is the number one problem within the wafer fabrication process (Turkot, 2017).

According to Smets et al. (2012), system availability is of crucial importance to produce cost-effectively. When a system failure occurs, and therefore the production stagnates, high cost can be considered (e.g. reduced production output). Since ASML has service level agreements with its customers, ASML takes a part of these costs on its behalf in the form of penalty costs.

Smets et al. (2012) further state that system availability is defined as the percentage of time a system operates *(uptime)* in a certain time interval. In addition, the time a system is not producing is called *downtime*. The uptime of the machine is quantified by the lifetime of the machine until a failure occurs. The time interval until a failure occurs is denoted as the time to failure with an average of Mean Time To Failure (MTTF). In literature, this is also known as the Mean Time Between Failure (MTBF) (O'conner & Kleyner, 2012). In addition, downtime is quantified by the time to repair, which is the time to bring the system back to its working condition (Thompson, 1999). The average time to repair is labeled as the Mean Time To Repair (MTTR). According to Smets et al. (2012), the time to repair depends on (1) the difficulty of disassembly or time to diagnose the problem (preparation), (2) the maintenance actions (do the actual job), and (3) the availability of spare/repair parts. ASML has a fourth variable, which is called recovery and implies the time to get a repaired machine back in its working condition.

For ASML, the availability of spare parts is especially important since some spare parts have a long lead time. Also, Hopp & Spearman (2001) note that waiting time for spare parts covers a large part of the repair time. We discuss the availability of spare parts in more detail in Section 2.3.

2.3 Spare part service management

In this section, we examine how ASML deals with the service levels of the delivered system as agreed in the SLA's with the customer. As mentioned in Section 2.2.3, the main focus of the research is on the probability distribution of time to repair and availability of spare parts. First, we describe the demand structure of spare parts in Section 2.3.1. Next, in Section 2.3.2, we discuss the supply of spare parts. Finally, we explain the SLAs and the different service measures in Section 2.3.3.

2.3.1 Demand structure of spare parts

In general, two reasons cause demand for spare parts. First, spare parts are required to perform preventive maintenance. This type of maintenance is performed in order to improve deteriorating machines and avoid failures. Since this type of maintenance is planned by forehand and, therefore, does not lead to unexpected downtime, we do not focus on this type of demand. The second reason for a spare part demand is appointed to corrective maintenance, which is performed when a machine breaks down. In the case a machine breakdown occurs, a customer contacts the ASML support centre, which sends an engineer. This engineer diagnoses the failure and requests a particular spare part from the local warehouse. When the spare part arrives, it replaces the broken part. The replacement of a broken part by a new spare part is known as the repair-by-replacement policy (Jiang, 2012). We display the demand for spare parts with the blue, orange and green arrows in Figure 2. The grey arrows represent the supply of spare parts. Most customers have SLAs with ASML, however, a few customers buy spare parts on occasion (Enders, 2004). Because these customers form a relatively small group, we do not take them into account in this research

2.3.2 Supply structure of spare parts

ASML has SLAs with its customers as mentioned in Section 2.1. As Jiang (2012) point out, ASML aims to find a trade-off between meeting the agreed service levels and minimizing both its capital investments (purchasing spare parts) and operational costs (holding spare parts). In order to meet the agreed service levels, ASML invested millions of euros in the last decade. According to Kranenburg (2006), a small proportion of cost reduction or reduction in spare part safety stocks leads to an absolute amount of saved money. We graphically display the stock points of spare parts from ASML in Figure 2. The structure consists of multiple Inventory locations including Global warehouse (GW), displayed in green, multiple continental/central warehouses (orange) and local warehouses (Blue). Since there are multiple stock points, the structure of spare parts at ASML is called a multi-echelon structure.

As mentioned in Section 2.3.1, an engineer orders a spare part at the nearest local warehouse. This local warehouse supplies this spare part if possible. If the needed spare part is locally not available, other nearly local warehouses are checked in order to supply this part. In literature, this is called lateral transshipments, see e.g. Ge, Peng, Van Houtum, & Adan (2018), and is especially beneficial (measured in time) when the central warehouse is located at a far distance (Enders, 2004). The part may also be supplied from the continental/central warehouse, which is called an emergency supply. Also, priority shipments are performed, which are slower and less expensive in comparison with an emergency supply. A spare part always passes the local warehouse before it arrives at the customer. When both the local warehouse and continental warehouse are not able to supply the needed spare part, a part is sent from the global warehouse. If a part is not in stock at this warehouse either, a part from one of ASML's factories is used to fulfill the demand. Obviously, this is the last and least preferable option (Enders, 2004). At the local warehouses, replenishment of the stock levels is done on a daily basis. Furthermore, demand is served according to a First Come First Serve (FCFS) policy.

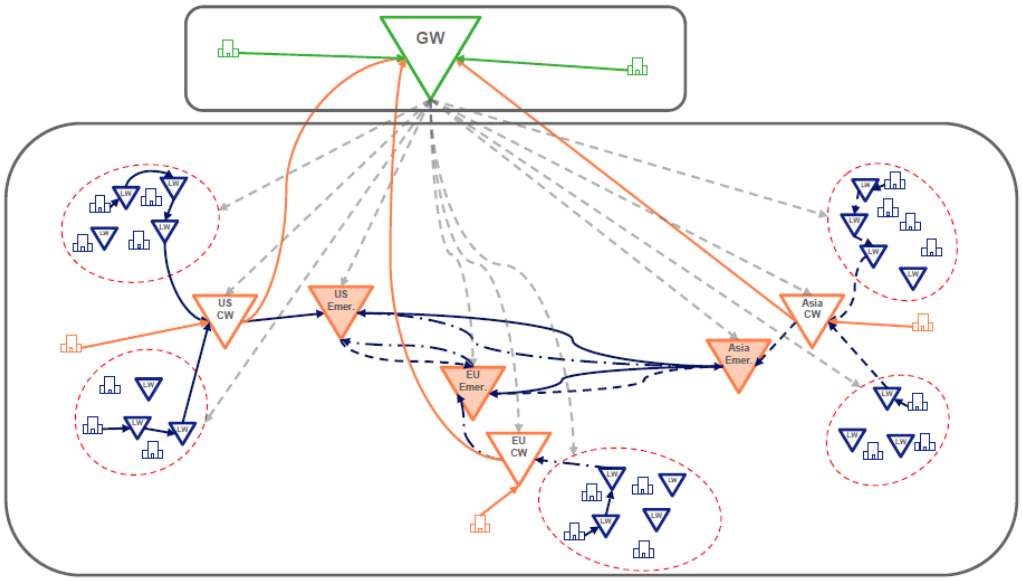


Figure 2 ASML service supply network (Van Aspert, 2014)

2.3.3 Service contracts

In this section, we first describe the different spare part service measures that are agreed in different service contracts of ASML. Then, we explain the role of the time horizon of these contracts. Service measures for spare parts are decision variables to control both the spare parts inventory and the system availability. A customer of ASML can choose certain measures (or combination) for spare parts, which will be included in the performance contracts.

Mean waiting time for a spare part

Downtime of a system consists of several components including, waiting for an engineer, the maintenance action and waiting for spare parts (Smets, et al., 2012; O'conner & Kleyner, 2012). Therefore, the mean waiting time for a spare part is only part of the total downtime. This is graphically displayed in Figure 3.

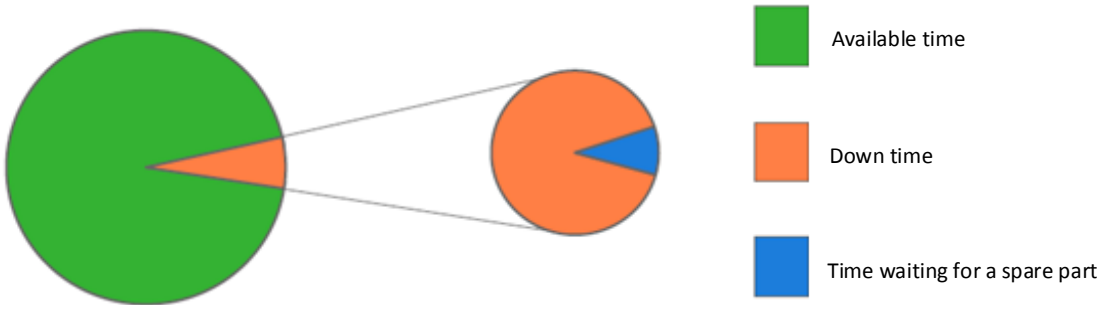


Figure 3 mean waiting time for a spare part (Van Aspert, 2014)

The mean waiting for a spare part service level indicates the expected time interval between the order of a spare part by an engineer until this spare part is delivered. This time interval varies due to the availability of a spare part and transportation time. The spare part is on stock at the nearest local warehouse and can be immediately delivered. Otherwise, the spare part is transported from another local warehouse (lateral transshipment), continental warehouse or even from the global warehouse (emergency transshipments). According to van Aspert (2014), the mean waiting for a spare part (in the formula set as DTWP) can be calculated with Equation 2.1. Within ASML, the following formula is the most used:

$$DTWP = \frac{Total\ system\ (s) unavailability\ in\ 13\ weeks\ due\ to\ transportation\ leadtime\ of\ parts\ (hours)}{24\ hours * 7\ days * 13\ weeks * \#\ of\ systems} * 100\% \quad (2.1)$$

As can be seen, the denominator includes 13 weeks since the mean waiting for a spare part is evaluated every quarter of the year. In this research, we focus on the steady state and therefore do not consider this time interval. The number of machines (systems) at the customer is taken into account. In literature, this service measure is called, expected waiting time or delay to fulfill a backorder (Basten & Van Houtum, 2014).

Local fill rate

This is the percentage of service parts that an OEM immediately can deliver from the local warehouse I the case that a customer orders a spare part. This means that, in contrast to the mean waiting time, the fill rate is not directly related to the emergency lead-time (Van Aspert, 2014). In literature, this service measure is known as the aggregate fill rate (Smets et al., 2012). According to Enders (2004), these fill rates are calculated for every individual customer and should be at least the percentage as established in the particular SLA. The fill rate is calculated with Equation 2.2.

$$fill\ rate = \frac{Demand\ supplied\ from\ local\ stock\ location\ of\ 13\ weeks}{Total\ demand\ of\ 13\ weeks} * 100\% \quad (2.2)$$

As well as the mean waiting for a spare part service measure is the fill rate calculated with the demand over a time interval of 13 weeks. Since we consider a long-term average (steady state), we do not take into account this time interval.

Extreme-long downtime (XLD)

This spare part service measure is a more direct performance measure, in contrast to the two earlier mentioned measures, since it is not based on averages, fractions or expectations. The measure is an agreed number of extreme long downs, longer than 12 consecutive hours, and therefore a fixed number. According to Lamghari-Idrissi et al. (2018), the XLD service measure provides more certainty enabling customers to better plan their operations.

Other spare parts service measures

In literature, there are more performance measures described including the number of backorders. This is the number of outstanding orders of spare parts. Another service measure, which is described in the literature, is the average availability. This service measure is closely related to the aggregate fill rate. However, instead of indicating the probability that a demand is fulfilled immediately, the average availability is the fraction of time that any given machine is available (Basten & Van Houtum, 2014). This is highly related to the repair time of machine and less on the waiting time for a spare part. Therefore, this service measure is out of scope.

Time horizon of service measurement contracts

Since service measures are agreed in contracts, the time horizon of this contract is important. In literature, finite horizon models are less researched in comparison with infinite horizon models. When a service provider based its planning on an infinite horizon, unnecessary costs are considered, e.g. holding costs are higher than necessary, higher transshipment costs due to emergency deliveries, and possible penalty costs. However, since ASML is planning on the steady state we assume the same and do not take contract durations in considering.

3 Research design

In this chapter, we describe the problem that we address in this research. In addition, we discuss a systematic approach in order to provide an answer to this problem. Based on the information from the problem context (Chapter 2) we define the following problem statement:

"Insufficient insight into the relationship between different spare parts service measures and the time to repair probability distribution leads to inefficiencies in the spare parts levels at ASML and lower results based of performance measures in the wafer fabrication process of the customer."

To find the consequences and causes of the problem, as given in the problem statement, a problem description is given in Section 3.1. Next, we discuss the scientific relevance of this problem in Section 3.2. Then, in Section 3.3, we provide the research questions and the methodology in order to develop a systematic approach to provide a solution to the problem statement. Finally, we describe the scope of the research in Section 3.4.

3.1 Problem description

Since the machine of ASML is one of the main bottlenecks in the production process of the customer, downtime of ASML's system results in high costs. These costs are mostly the result of the loss of production and are more than 20 euros per second or 1.5 million euros per day (ASML, 2018b). A fab manager of one of the main customers of ASML states that downtime of an ASML system results in a standstill of 20% of the entire production line (Hasan, 2016). This standstill is among other things due to starvation. Kiesmüller & Zimmerman (2018 p. 369) describe starvation as follows: "when a WIP buffer before a machine is exhausted, and there are no other WIP parts available, the machine cannot produce and consequently stands still". In order to respond to the downtimes of the bottleneck machine, customers increase the WIP buffers in the production system (Hopp & Spearman, 2001). However, this is costly and results in deterioration of the system's performance measures. To find the optimal WIP level of a system, customers require an accurate prediction of the bottleneck downtime and, therefore, the repair lead time. In the case that ASML provides this information, higher customer satisfaction can be achieved.

As noted in Section 2.3.2, ASML aims to find a trade-off between meeting the agreed service levels (spare parts availability) and minimizing the costs of holding spare parts and transportation costs. So, on one hand, the customer aims to reduce the WIP level within the fabrication process and on the other hand, ASML aims to reduce the spare part inventory level under the condition that agreed SLAs are satisfied. This compromise is called, *double buffering*. However, the WIP level also depends on the other two performance measures (i.e. cycle time and throughput). In literature, the trade-off for ASML and the performance measures are addressed into two parts, which we describe in the following paragraphs. First, the relationship between spare parts and time to repair probability distribution. Second, the relationship between the coefficient of variation of the time to repair and the performance measures of a production system.

Spare parts service measures in the literature

According to Basten & Van Houtum (2014), the minimal spare part holding costs can be found under the constraints of different spare part service measures. More spare parts than the optimum leads to higher holding costs and a shortage of spare parts leads to higher penalty costs. Furthermore, a spare part service

measure can be fitted into a time to repair probability distribution, which is often a lognormal distribution (Smets, et al., 2012; O'conner & Kleyner, 2012).

The coefficient of variation of the time to repair in literature

We follow Hopp & Spearman (2001) to find the relationship between the time to repair probability distribution and the WIP level of a production system. Therefore, we define σ_r as the standard deviation and c_r as the coefficient of variation (CV) of the repair times. Furthermore, we determine t_0 and c_0 as the natural processing time and the coefficient of variation in the natural process, respectively. To calculate the CV of the effective process time (c_e), we use the following equations:

$$c_r = \frac{\sigma_r}{MTTR} \tag{3.1}$$

$$c_e^2 = c_0^2 + A(1-A)\frac{MTTR}{t_0} + c_r^2 A(1-A)\frac{MTTR}{t_0} \tag{3.2}$$

Note that the first term represents the variability of the natural processing time, the second term the random outages of the system, and the third term is due to the variability of the repair time. An increase in the coefficient of the repair time leads to an increase in the coefficient of the effective process time. Furthermore, it is notable that the second and third terms are increasing in the mean time to repair. This indicates that longer repair lead times cause higher variability as long as all other variables remain equal. So, a machine with frequent short outages is preferred over one with infrequent long outages, as long as the system availability is the same. Gupta et al. (2006) confirm this statement and state that high variability at a system leads to less accurate predictions of the production cycle time. Moreover, the variability at one station (especially when this is the bottleneck station) affects the behavior of the whole production line, which is called flow variability. Variability in an upstream machine often affects the machines more downstream in the process. Finally, a higher variation of effective process time results in lower throughput, longer cycle time and more WIP inventory.

Conclusion

In summary, a higher variation in the spare parts repair lead time results in a higher variation of the repair lead time. A higher variation of the repair lead time results into a higher variation of effective process time. This leads to more starvation in the production process, which causes a deterioration of performance measures within the manufacturing process. The deterioration of the performance measures means a higher WIP level, longer cycle time and less throughput.

3.2 Scientific relevance

According to Van Aken et al. (2012), the relevance of a business research project is based on both operational relevance (Section 3.1) and scientific relevance, which we provide in this section. To report the scientific relevance, we first describe the research that is done at ASML based on spare parts service measures. Then, we discuss the current research gaps based on spare part service measures and the relationship between the time to repair probability distribution and performance measures of a production system. Finally, we provide the gaps that we will actually address in this research.

Research at ASML

According to Enders (2004), there is a close research-based collaboration between ASML and Eindhoven University of Technology (TU/e) in the field of service management. Within this collaboration, Kanters (1995) studies a metric approach in order to improve inventory control, as described by Sherbrooke (1992). Furthermore, Kranenburg (2006) studied both customer differentiation and the incorporation of lateral supply in the supply chain planning. Within the research-based collaboration, the role of the central warehouse is a popular subject. Jiang (2012) investigates the role of the central warehouse in the spare parts inventory control and Van Aspert (2014) examines a design of an integrated global warehouse and field stock planning concept for spare parts. All the previously mentioned studies focused on the relationship between the availability of spare parts and system availability. However, this is inadequate for measuring the influence of spare part availability on the performance measures of a production process.

Research gaps based on spare parts service measures

We identified three different literature gaps based on double buffering while concerning spare parts service measures. Firstly, researchers mainly discussed service measures that provide an average or expected availability instead of a distribution of the availability. In addition, the relationship with the time to repair probability distribution is even less researched. Secondly, time restricted spare part service measures are rarely explained in literature such as extreme-long down times. Only Lamghari-Idrissi, et al (2018) discuss three heuristics to find the optimal base stock levels under these service measures. However, these heuristics are only applicable in a single stock, single item model. A research gap is located by extending these heuristics to a multi-item, multi-stock model. Thirdly, spare part service measures are often determined on an infinite horizon, which leads to higher risks of not meeting the agreed level (Lamghari-Idrissi et al., 2018). Research on these three research gaps will be beneficial for the semiconductor supply chain since service measures will lead to more predictability of the time to repair probability distribution and thus system availability.

Research gaps based on the influence of the time to repair distribution and performance measures

Only Kiesmüller & Zimmerman (2018) investigate the relationship between spare parts provisioning and a performance measure of a production system. Because these authors were the first who describe this relationship, a relatively simple model is used. Furthermore, the aim of the paper is to be a building block for further research based on this topic. A literature gap can be filled by investigating the influence of spare parts on the performance measures of a more complex system. Addressing this gap is beneficial for processes where costs rise exponentially in the case that a system is down such as the airline industry. When the waiting time probability distribution for spare parts is known, manufacturers are better able to predict the downtime. Therefore, better results based on the process performance measures are achieved.

Research gaps this research addresses

Investigating all the noted gaps is too extended to touch in this research. Therefore, we only consider the gaps that affect ASML the most, which are the following:

- The influence of spare part service measures on the time to repair probability distribution,
- The influence of the time to repair probability distribution (coefficient of variation) on the performance measures of a wafer fabrication process (production process with recirculating flows).

3.3 Research questions and methodology

In this section, we describe the assignment and approach of the research. In the problem statement, we have mentioned that the relationship between spare part service measures and the performance measures of a wafer fabrication is unclear. Based on this information we define the following main research question:

Main Research question:

"How do different spare part service measures and their parameters influence the performance measures of a wafer fabrication process?"

To answer the main research question, we answer multiple research questions within two parts, which we graphically display in Figure 4. Within this figure, the mean waiting time for a spare part service measure is denoted as DTWP as it is called as ASML. In Part I, we focus on the research topic: service management. Then in Part II, we treat the research topic: operations management.

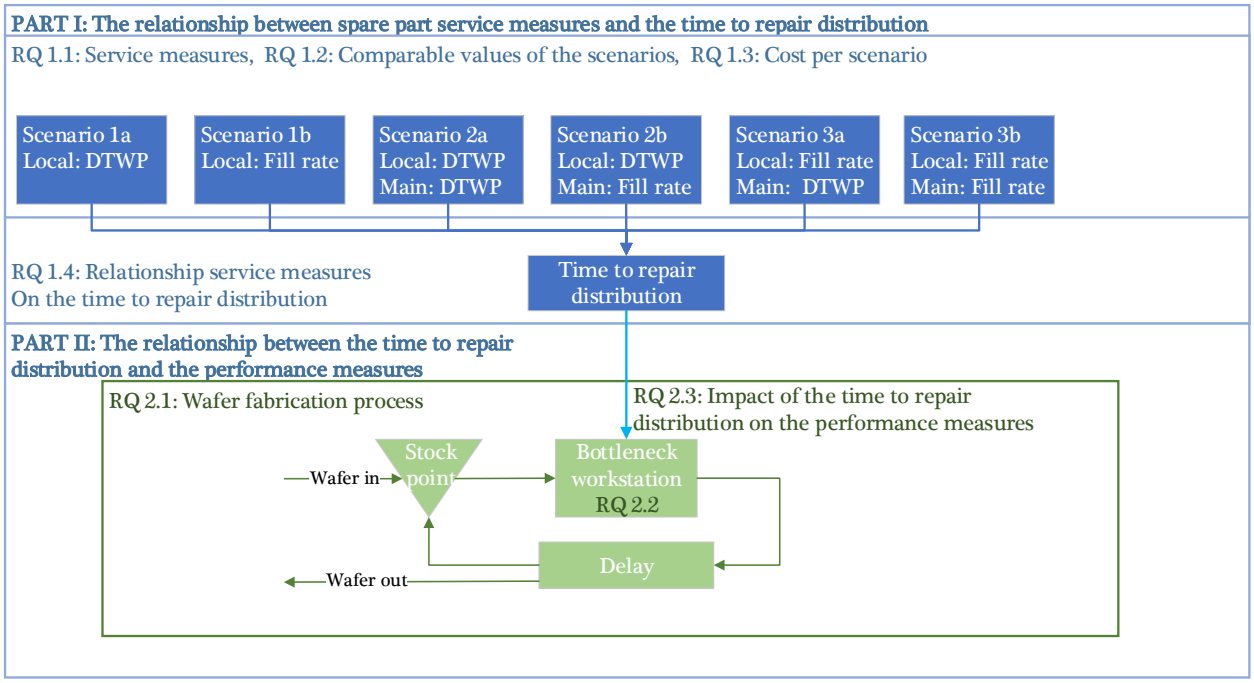


Figure 4 Relationship between spare part service measures and wafer fabrication process

First, we investigate Part I, where we focus on the relationship between the service measures and the time to repair probability distribution. This leads to the following research question:

Research Question of Part I:

"How do different spare part service measures and their parameters influence the total inventory costs and the probability distribution of the time to repair?"

Figure 4 shows that there are six different scenarios, which may have a different effect on the time to repair probability distribution. A customer can determine which spare part service measure or combination (i.e. fill rate and mean waiting time for a spare part) is included in the SLA. In the scenarios 1a and 1b, we only focus on a single location model. However, service measures are also influenced by the supply chain network. Therefore, we also investigate a scenario with an extra warehouse, which is called the main warehouse. This main warehouse is able to send lateral transshipments and has a customer with a service contract as well.

In order to investigate the different variables that influence the relationship between the spare part service measures and the time to repair, we start with the following research question:

RQ 1.1 Which variables influence the different spare part service measures and how to model these?
We investigate the service measures by discussions with employees of the service management department. Furthermore, we use the following literature; Reijnen (2009), Van Houtum & Kranenburg (2015) and Aspert (2014) to derive a mathematical model. We especially focus on the spare part lead time in practice since it has a huge impact on the service measures (Houtum & Kranenburg, 2015). We analyse this and other variables using historical data. In order to compare the different spare part service measures, we calculate the corresponding values that result in the same mean waiting time for a spare part. This leads to the following research question:

RQ 1.2 What are the comparable values of the different spare part service measures?
To compare the different spare part service measures, we use the mathematical model as described in the first research question. Also, we use ASML's tool called SPartAn, which is used to calculate the optimal base stock levels by focussing on different spare part service measures. Using this tool, we can find commensurate values of the different scenarios and the necessary base stock levels. The next step is to calculate the costs for the spare parts levels corresponding to the service measures, which leads to the next research question:

RQ 1.3 What are the holding costs of spare part inventories based on the different corresponding service measures for spare parts?
When all the corresponding service measure values are provided, the spare parts holding costs will be investigated for every scenario. We will investigate the costs based on data from SAP and interviews with employees of the finance department. Also, the literature based on inventory costs within companies will be used such as Nahmias and Olsen (2015) and Datta & Roy (2009). Finally, we analyse the influence of the scenarios on the probability distribution of the time to repair, which leads to the following research question:

RQ 1.4 What are the probability distributions of the time to repair that results from the different scenarios?
According to Jiang (2012), there is a relationship between the different service measures and the probability distribution of the time to repair. As noted in Section 2.2.3, the time to repair consists of multiple components where we only consider the waiting time on a part as stochastic. The other components are assumed as deterministic since we focus on the spare part service measures. In order to provide a different time to repair probability distributions, we develop a simulation model. This simulation model is based on the mathematical model of RQ 1.1 and uses the base stock levels of RQ 1.2. We validate the simulation model based on SPartAn, by visual inspection and by discussions with employees of the service management department.

Part two is based on the wafer fabrication process of the customer of ASML. This contains the wafer fabrication process as described by Akcalt, et al (2001), May & Spanos, (2006) and Mönch, et al. (2017). To investigate this process, we provide the following research question:

Research Question of Part II:

"How do different time to repair probability distributions, resulting from different spare part service measures, influence the performance measures of a wafer fabrication process?"

We start the second part by analysing the fabrication process with the research question:

RQ 2.1 What are the characteristics of the wafer fabrication process?
According to Dijkman (2017), multiple discovery techniques, including interviews and data analysis, can be applied to analyse different process steps. To obtain a clear vision of the wafer fabrication process, we will take interviews with the infab logistics department. These employees have information about the production line and the role of the ASML system within this line. Furthermore, Akcalt et al. (2001) provide a scaled-down representation of a photolithography process with different production steps, routings, process times, and setup times. This information will be validated by interviews with employees of the customer service department at ASML. Within this research question, we especially focus on the following elements:

- Process steps and routing
- Bottleneck processes
- Performance measures

However, since the production process of wafers is very complicated, numerous authors (including Upsani et al., 2006) only focus on the process steps that require the bottleneck workstations. The remaining workstations are modeled by delays in form of waiting and processing times (Mönch, et al., 2017). Therefore, we will roughly describe the wafer process in a business process management notation (BPMN) (Dijkman, 2017) and then describe the bottleneck process in more detail. This leads to the following research question:

RQ 2.2 What are the characteristics of the bottleneck process in the wafer fabrication process?
The bottleneck process of the wafer production process is described by multiple authors. We will use different literature, i.e. May & Spanos (2006), Mönch et al. (2017), Akcalt, Nemoto, & Uzsoy (2001), to investigate the bottleneck. Furthermore, we will use historical data obtained from SAP to investigate the time to failure of the machines. The results will be validated by interviews with employees of the department service management at ASML. To answer this research question, special attention is given to the following elements:

- Processing time
- Time to failure
- Dispatching rules
- Utilization

When the bottleneck process is analysed, we focus on the total wafer fabrication process, shown in Figure 4. This leads to the following research question:

RQ 2.3 What is the impact of the different repair time probability distributions, corresponding with the multiple scenarios, on the performance measures of a fabrication process?
To answer this research question, we use Hopp & Spearman (2001), who discuss multiple formulas to investigate the relationship between the repair time and variance of the bottleneck and the performance measures of a system. Furthermore, the papers of Kiesmüller & Zimmerman (2018) and Lin and Lee (2001) will be used as a guideline to design a simulation model.

We will analyse the production system by simulation since this is the most common modeling in the semiconductor industry (Lin & Lee, 2001). To be more specific, discrete-event simulation is noted to be a well-established tool for analysing wafer fabs (Mönch et al., 2017). Furthermore, by using the research of Law (2007), a technically correct simulation can be developed with all important aspects of a simulation study such as warm-up time and the number of simulation runs. Finally, the simulation model compares the impact of the six repair time probability distributions on the performance measures of the wafer fabrication process in order to provide an answer to the main research question.

3.4 The scope of the research

To define the project and clearly indicate its boundaries, we draft the scope per part of the project:

Part I

- The time to repair exists of multiple causes, where we only model the waiting on spare part-time as stochastic. This is due to the fact that we are only interested in the impact of the service measures and thus only the time waiting for a spare part. In order to describe the relationship as clear as possible, we assume that other causes are deterministic e.g. preparation, maintenance action and preparation time.
- We only consider the NXT machines since there is a lot of historical data available about these machines.
- Within the supply chain network of spare parts, we only consider two models. On the one hand a model with one global warehouse and a local warehouse. On the other hand, a model with one global warehouse, a local warehouse, and the main warehouse. We choose to only consider this part of the spare part supply chain in order to show the difference between a single and a multi-stock model. We only provide two local warehouses, besides the global/central warehouse, since we expect no new insights by extending this network.
- We only consider demand from aftersales events i.e. corrective maintenance. Demand from installing, upgrading or replacing a system is out of scope since these events are planned.
- We only analyse the service measures that are actually offered by ASML. (i.e. mean waiting time for a spare part and fill rate). The XLD service measures are out of scope since there is no multi-item, multi-stock optimization model or heuristic (noted in Section 3.2) to determine the optimal or near optimal base-stock levels. Investigating this subject is a research in itself and too extended to take part in this research.

Part II

- Within this research, we only focus on the wafer fabrication process. We do not take into account the rest of the wafer manufacturing, such as electrically probing, assembling and testing of the wafer because ASML machines are not used in these processes.
- In literature, the time to support, which is the time between a system down is reported and an engineer arrives at the customer, is also pointed out as a cause of downtime (Smets et al., 2012). We consider the time to repair and not examine the time to support since engineers are on-site.
- The focus within the wafer fabrication process is on the bottleneck machine since this machine is provided by ASML. In addition, in the wafer fabrication process, the bottleneck has significantly more impact on the performance measures since the process consists of multiple loops. Other workstations will be modeled as a delay, so not modeled in much detail. To model the nonbottleneck processes as a delay is common in research about wafer fabrication (Monch et al., 2017a).
- We only model critical components. This implies that we only consider the components that, when failing, directly results in downtime of a machine. Furthermore, we note that a critical component can be replaced by a single spare part (SKU).
- In the wafer fabrication simulation model, degradation of the work station's output is not taken into account since this information is not available.
- We do not take into account the technical state of a bottleneck machine. So, regardless of the production years, the machine output remains the same.
- Finally, we do not focus on part swapping or quick fixes to get the machine operating again. If a part failed, it is replaced by a spare part and the machine gets back in its working condition. We do not consider other maintenance actions since they are not related to the spare part service measures.

4 The influence of spare part service measures on the probability distribution of the time to repair

This chapter contains the first part of the research, answering the first research question:

"How do different spare part service measures and their parameters influence the total inventory costs and the probability distribution of the time to repair?"

As mentioned in Section 2.2.3, we divide the time to repair into the components; preparation, waiting time for a spare part, maintenance action, and recovery. In order to investigate the effect of spare part service measures, we focus on the waiting time for a spare part. The other three components are assumed to be deterministic. The time waiting for a part used to be a small part of the time to repair of NXT machines on average. However, the waiting time for a spare part was as second largest component responsible for the variation within the time to repair. Both are based on the engineers NXT machine failure data of 2017 and we provide those in order to give an indication of the time to repair components. We validate this information with Employee 1 (See Appendix B for the department and the title of the referred employees).

We start this chapter by describing the spare parts mathematical model in order to explain the spare part supply chain network, inventory costs, spare part service measures and the inventory levels in Section 4.1. Then, in Section 4.2, we analyse the supply chain lead time probability distribution of ASML since this is of great importance for calculating the service measures. Thereafter, we develop a simulation model where we use the calculated inventory levels and analysed supply lead time probability distribution of ASML as input. The goal of the simulation model is to provide the probability distributions of the waiting time of a spare part by applying different spare part service measures. In addition, we describe the transition from these distributions into time to repair probability distributions. We provide the simulation and the probability distributions in Section 4.3. Finally, we discuss the conclusions in Section 4.4.

4.1 Spare parts mathematical model description

In this section, we provide a general explanation of the investigated model. The goal of this section is to (i) describe the investigated supply network through mathematical expressions, (ii) show how the different service measures can be calculated and are related to each other, (iii) provide formulas to calculate the total inventory costs and, (iv) explain an optimization problem to calculate the base stock levels corresponding to the different service measures. Hereby, we follow Van Aspert (2014), who describes a mathematical model for the whole spare part supply chain of ASML. Since we investigate a scaled-down representation of the total spare part supply chain of ASML, we do not take the entire model of the author into account.

In this research, we focus on NXT machines, which consists of multiple critical parts. These parts can be replaced with a spare part when a failure occurs and are defined as stock keeping units (SKUs), which are held at a warehouse. We denote the set of different SKUs as $I = \{1,2, \dots, |I|\}$. These different SKUs are used in the mathematical model that consists of four parts (i.e. the supply network, costs, service measures and stock levels). We first describe the supply network of the model in Section 4.1.1. Then, in Section 4.1.2, we explain the cost function based on stock levels and transportation costs. Then, we appoint the formulas to calculate the service measures in Section 4.1.3. Finally, we describe the optimization model to determine the stock levels based on the network, costs, and service measures objectives (i.e. mean waiting time for a spare part and fill rate) and provide the main assumptions in Section 4.1.4 and 4.1.5, respectively.

4.1.1 The supply network

As noted in Section 2.3.2, the network of ASML consists of a global warehouse, multiple continental, and local warehouses. Within the total network of ASML, we only focus on two versions, see Figure 5. On the one hand, a single inventory model with only one local warehouse (LW) and, on the other hand, a network model with an additional main warehouse. The main warehouse (MW) is similar to a local warehouse and provides customers of spare parts. However, the main warehouse is, in contrast to the local warehouse, able to send lateral transshipments to the local warehouse that is responsible for plan group n. We assume that the emergency and replenishment supply of spare parts are always delivered from the central warehouse (CW). In addition, the lead times for an emergency or replenishment supply are equal for both the local and the main warehouse. Otherwise, it could be faster to send a spare part to the main warehouse and then with lateral transshipment to the local warehouse instead of an immediately delivery from the central warehouse to the local warehouse. Moreover, we assume ample stock for the central warehouse. So, if there is an emergency supply needed, the central warehouse is always able to send the requested part.

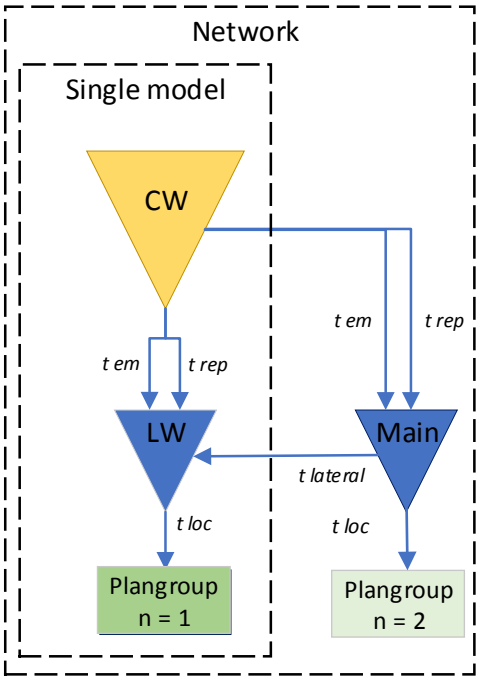


Figure 5 Investigated supply models

We describe the model based on the most extensive supply chain that we investigate, which is the network as given in Figure 5. Within the network, we let $J = \{1, 2\}$ be the set of warehouses where SKUs are kept on the stock. Each warehouse uses a base stock level (*S-1, S*) policy and delivers by a first come first serve procedure. We assume this policy since ASML currently applies this policy as well. The base stock level of SKU i at warehouse j is given by $S_{i,j}$, while $\boldsymbol{S}$ denotes the matrix of all base stock level that is defined as:

$$\boldsymbol{S} = \begin{pmatrix} S_{1,1} & S_{1,2} \\ \vdots & \vdots \\ S_{|I|,1} & S_{|I|,2} \end{pmatrix}$$

As noted in Section 2.3.2, customers are classified into geographical areas (e.g. Japan, EU, Korea, US). When several customers are located close to each other, they form a plan group. We denote the set of plan groups by $N = \{1, 2\}$, where each plan group has a designated warehouse $j \in J$.

In our model, we consider two warehouses, which are both responsible for a plan group. For plan group $n = 1$ the array $v_1 = (1,2)$. So, a demand is first satisfied from local warehouse 1 if this warehouse has the part on stock. Otherwise, the main warehouse ($j = 2$) is checked if this part is on stock. If both warehouses cannot satisfy the demand, the part is sent from the global warehouse as an emergency order We assume that the global warehouse can always send an emergency shipment. Furthermore, we denote for plan group $n = 2$ the array $v_2 = (2)$.

We denote the number of machines within a plan group by Z_n. The total demand of SKU i from all machines together in plan group n, is given by the fail rate ($m_{i,n}$). In addition, we denote the total demand of all machines at the plan group n as $M_n =$. The demand stream per SKU is assumed to be an independent Poisson process. Moreover, we assume that all lead times are exponentially distributed in since in this way the model can be analytically solved, see e.g. Van Houtum & Kranenburg (2015). We identify the following transportation times, which we all assume to be independent and identically distributed:

- t_i^{loc} = Average lead time for a local shipment for SKU $i \in I$
- t_i^{em} = Average lead time for an emergency shipment to a local or main warehouse for SKU $i \in I$.
- t_j^{rep} = Average replenishment lead time for local or main warehouse $j \in J$.

With respect to the demand fulfillment, we use the following notation:

- $\beta_{i,n,j}(S_{i,j})$ as the probability that positive stock on hand at warehouse j is observed when a demand form plan group n of SKU i occurs.
- $\theta_{i,n}(S_{i,j})$ as the probability that a demand for SKU i by plan group n is not fulfilled by any of the warehouses (not considering the global warehouse).

The following expression is used to calculate the latter fraction for plan group $n = 1$:

$$\theta_{i,1}(S_{i,j}) = \left(1 - \beta_{i,1,1}(S_{i,1})\right) * \left(1 - \beta_{i,1,2}(S_{i,2})\right) \quad (4.1)$$

For plan group $n = 2$, we denote the following formula:

$$\theta_{i,2}(S_{i,j}) = \left(1 - \beta_{i,2,2}(S_{i,2})\right) \quad (4.2)$$

4.1.2 The costs

The costs of the spare parts supply chain consist of the components: total holding costs (C_i^{h}), the costs for lateral (C_i^{lat}) transhipments and emergency (C_i^{em}) shipments for all SKUs $i \in I$. We do not consider both the costs for replenishment orders and the cost for shipping parts from the local warehouse to the plan group since these are the same for all the different scenarios. We calculate the total costs using the following formula:

$$C(\boldsymbol{S}) = \sum_{j \in J} \sum_{i \in I} C_i^{\mathrm{h}}(S_{i,j}) + C_i^{\mathrm{lat}}(S_{i,1}, S_{i,2}) + C_i^{\mathrm{em}}(S_{i,j}) \quad (4.3)$$

We explain the three different components of the total costs per unit in three different paragraphs. We start with the holding costs, then discuss the lateral transportation costs and, finally, provide the emergency transportation costs.

Holding costs

With respect to the holding cost, we follow Nahmias and Olsen (2015, p. 204-205). The authors state that holding costs, also known as carrying costs or inventory costs, is defined as: "the costs that accrue as a result of having capital tied up in inventory". In other words, holding costs are the sum of all costs related to inventory of a process or supply chain. Within the research, we do not consider holding costs for SKUs that are in transit. Furthermore, the authors explain that holding costs consist of the following components:

- Costs of the physical space to store a stock keeping unit (rental costs)
- Taxes and insurance
- Obsolescence, breakage, spoilage, and deterioration
- Opportunity costs

Opportunity costs are defined as the value that could be obtained when the investment for these costs were invested in an alternative. Thus, decreasing inventory levels results in an increase of capital, which could be invested internally or externally. The total percentage of holding carrying costs is given by the sum of the four components as given above expressed in the holding cost rate (R). In general, we measure the inventory in units and express the holding cost in terms of euros per unit per year. This results in the following formula:

$$c_i^{\text{stock}} = R * c_i^{\text{v}} \tag{4.4}$$

With,

c_i^{stock} = Stocking cost in euros per unit per year of SKU $i \in I$;
c_i^{v} = The value of one unit of SKU $i \in I$.

We assume that all elements are the same for each warehouse. The total holding costs for SKU i is given by:

$$C_i^{\text{h}}(S_{i,j}) = c_i^{\text{stock}} * S_{i,j} \tag{4.5}$$

Lateral transshipment costs

Lateral transportation costs occur when a lateral transshipment is performed. In the researched model, only demand of plan group 1 $(n = 1)$ can lead to lateral transshipments. Costs for lateral transshipments occur if the local warehouse is not able to satisfy demand and the main warehouse is able to do so. We denote the costs for lateral transshipments as c_i^{lat} and provide the following equation to calculate the lateral transhipment costs:

$$C_i^{\text{lat}}(S_{i,1}, S_{i,2}) = m_{i,1} * \left(1 - \beta_{i,1}(s_{i,1})\right) * c_i^{\text{lat}} \beta_{i,2}(s_{i,2}) \tag{4.6}$$

Emergency transshipment costs

We follow Van Houtum & Kranenburg (2015), who point out that the total cost of emergency transshipments for SKU $i \in I$ is equal to:

$$C_i^{\text{em}}(S_{i,j}) = \sum_{n \in N} m_{i,n} * c_i^{\text{em}} \theta_{i,n}(S_{i,j}) \tag{4.7}$$

Whereby c_i^{em} is the cost for sending an emergency transhipment for SKU i from the central warehouse to a local warehouse. Since one less replenishment transhipment to the stock point is required when an emergency transhipment is performed, these replenishment transhipment costs must be subtracted. However, we do not take into account replenishment transhipment costs since these are the same for each service measure.

4.1.3 Spare part service measures

We divide this section into the different scenarios as mentioned in Section 3.3. Firstly, we describe the fill rate service measure. Then, we explain the mean waiting time for a spare part service measure. Finally, we conclude the differences between the service measures and explain the different scenarios, which we investigate.

Since both the service measures are highly dependent on the characteristics of the supply network, we divide this scenario into two. As noted in Section 4.2.1, there is a difference between a single stock model and a multiple stock model. In this latter scenario, lateral transshipments are applied. However, the number of lateral transshipments depends on the sort of service level agreement of the main warehouse. Therefore, we investigate the difference between a mean waiting time for a spare part and a fill rate agreement for the main warehouse.

Fill rate

In literature, this service measure is actually called the *Aggregate Fill Rate* and is the fraction of demand that is immediately satisfied from the designated warehouse. Note that the fill rate is the average of the satisfied demand of a warehouse so some parts have a higher availability than others. For example, we set a fill rate of 95%. This does not mean that the demand of each individual part is satisfied with 95% of the cases. The aggregate fill rate implies that 95% of the total demand must be satisfied from the warehouse. Furthermore, Van Houtum & Kranenburg (2015) provide the following formula to calculate the Aggregate Fill rate:

$$\beta_j(\boldsymbol{S_j}) = \sum_{i \in I} \frac{m_i}{M} \beta_{i,j}(s_{i,j}) \tag{4.8}$$

Hereby, the fill rate for SKU i, also known as the *item fill rate*, is denoted by $\beta_{i,j}(s_{i,j})$. This fill rate is the probability that a positive stock on hand is observed when a demand of SKU i occurs. Under the application of emergency shipments an M/G/c/c queue can be applied, which is also known as an *Erlang loss system*. In other words, there are $c = S_i$ parallel servers, an arrival rate of m_i and mean service time of t_j. Since the fill rate is the fraction of time that there is at least one part on stock, this is equal to the fraction of time that at least one server is free. So, this can be calculated by one minus the fraction of time that all servers are occupied. Thus,

$$\beta_{i,j}(s_{i,j}) = 1 - \frac{\frac{1}{s_{i,j}!}\rho_{i,j}^{s_{i,j}}}{\sum_{k=0}^{S_{i,j}} \frac{1}{k!}\rho_{i,j}^{k}} \tag{4.9}$$

Where, $\rho_{i,j} = m_i t_j$.

Mean waiting time for a spare part

The mean waiting time for a spare part service measure is the average fraction of time that a system is down and is waiting on a spare part. So, in contrast to the fill rate service measure, the mean waiting time for a spare part measure has a time restriction. To calculate the mean waiting time for a spare part service measure, we start with calculating the mean waiting time of a delivery of SKU i for plan group n:

$$W_{i,n}(S_{i,n}) = t_i^{\text{em}}\theta_{i,n}(S_i) + \sum_{j\in J} t_{n,j}^{\text{loc}}\beta_{i,n,j}(S_{i,j}) \tag{4.10}$$

Note that $\theta_{i,n}(S_i)$ is given in Equation 4.1. Furthermore, the mean waiting time for a spare part service measure contains all the warehouses that can satisfy the demand of plan group n. This is different from the fill rate service measure that only considers the designated warehouse. To analyse the time that a certain SKU i is unavailable, we multiply the expected waiting time by the expected demand ($m_{i,n}$) per time unit for SKU i. This demand is also known as the failure rate of SKU i. Now, we can provide the following equation for the total waiting time for a spare part:

$$W_{i,n}(S_{i,n}) * m_{i,n} \tag{4.11}$$

If we divide the unavailability by the number of machines at plan group n, which we denote as Z_n, the mean waiting time for a spare part (In the formula denoted as DTWP) for SKU $i \in I$ over the time interval t^{int} is calculated. This results in the following formula:

$$DTWP_{i,n}(S_{i,n}) = \frac{W_{i,n}(S_i) * m_{i,n}}{Z_n * t^{int}} \tag{4.12}$$

Finally, we provide the mean waiting time for a spare part of all SKUs for plan group n with:

$$DTWP_n(\boldsymbol{S_n}) = \sum_{i\in I} DTWP_{i,n}(S_i) \tag{4.13}$$

Different scenarios

The main difference between the fill rate and the mean waiting time for a spare part measures is the time constraint. Van Houtum & Kranenburg (2015) assume that an order that is not satisfied by the local warehouse, is always fulfilled as soon as possible. It is important to note that this is also done by ASML because of the focus on high customer satisfaction. However, theoretically, this does not always hold. For example, when demand of a certain SKU occurs, a local warehouse is out of stock and the continental warehouse has only one part on stock. This part can be kept on stock for a more important plan group. So, the particular part will be delivered from the global warehouse, which obviously takes more time. Especially, in the case that a local warehouse misses an order under a fill rate agreement, there is no time constraint for delivering this missed order. In other words, under the fill rate agreement, ASML has not the responsibility to deliver the part as soon as possible. Therefore, we assume that the emergency time of a fill rate agreement is equal to the replenishment time.

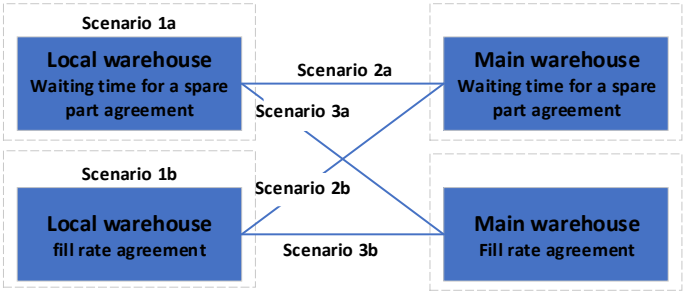


Figure 6 The different scenarios

In this research, we investigate multiple scenarios as shown in Figure 6. The first two scenarios (1a and 1b) are based on a single stock point location. Next, scenario 2a and 2b are characterized by a network whereby the main warehouse has a mean waiting time for a spare part agreement. Additionally, the last two scenarios (3a and 3b) are focused on a network as well; however, the main warehouse has a fill rate agreement. Furthermore, the scenarios that contain an 'a' have a normal emergency time in contrast to the 'b' scenarios, which contains an emergency lead time that is equal to the replenishment lead time. This is due to the fact that in theory the fill rate agreements are not restricted by time.

4.1.4 The inventory levels

In the mathematical model, the base stock level $S_{i,j}$ (mentioned in Section 4.1.1) is a decision variable. Since we investigate two different service measures, we consider the same number of optimization problems with the goal (objective) to minimize costs. The restrictions of the optimisation problems are different because the mean waiting time for a spare part is time restricted and the fill rate is restricted on the fraction of spare parts that is immediately satisfied from the local warehouse. Both problems are considered by Van Aspert (2014), which we provide in Appendix C. For the optimization procedures, we refer to this author as well.

4.1.5 Summary of the assumptions about the service measures

In this section we briefly summarize the most important assumptions, which are:

- Only a single stock model and a spare part supply chain model with two warehouses are investigated.
- Only the main warehouse is able to send lateral transshipments.
- Both warehouses deliver by First Come First Serve and have a base stock policy.

4.2 The supply lead times in practice

Next to the theoretical perspective, as described in Section 4.1, we analyse the service measure model in practice as well. Since the supply lead time has a crucial role on the results of the service measures, we focus explicitly on this probability distribution. To analyse supply lead time probability distribution of ASML, we follow Banks et al. (2001). These authors state that an input data model is developed in four steps, which are: (i) gathering data from a real-world system, (ii) identify the statistical distribution using a histogram, (iii) identify the corresponding parameters and (iv) evaluate the statistic distribution by using a statistical test (i.e. Chi-square) to determine the goodness of fit. The described steps correspond to the titles of the paragraphs in this section.

Gathering information

We start by gathering the historical data from the End-to-End database of ASML to investigate the three different supply lead times (i.e. local, lateral and emergency lead time) as noted in Section 2.3.2. The database provides transportation data from July 2016 until January 2018. Within the database, transportation time is denoted as the time since a demand occurs at the customer until the customer receives the spare part. Based on the spare part supply chain, we focus on the warehouses that support the largest customer with a mean waiting time for a spare part agreement and the largest customer with a fill rate agreement since:

- These warehouses provide components for more than 15 NXT machines.
- These warehouses have comparable local and emergency lead times.
- These warehouses are a good reflection of all warehouses in the field.

We validate the reliability of this data by Employees 6, 7 and 11.

Identify the statistical distribution and the corresponding parameters

Based on the data of the End-to-End lead time database, we calculate the mean and variances for the different lead times (see Table 2). To find the probability distributions that fit the data best, we use an distribution fitting application of Matlab. This application transforms the historical data into histograms where we use the Diaconis rule of Freedman & Diaconis (1981) to determine the bin width. Then, we analyse all common distributions including exponential, gamma, lognormal, Weibull and the normal distribution, which we all plot on one histogram. Hereby, we calculate the best parameters to scale the distributions to the best fit by using the application. We provide the best-fitted distribution by visual inspection and a comparison between distributions is graphically displayed in Appendix D. We conclude that all three data sets best fit a lognormal distribution. This is reasonable in practice since more parts arrive later rather than earlier as compared to the modus (De Treville et al., 2014). However, the fit of the emergency lead time seems unsatisfactory, unlike the other two lead times, since it does not fit well at the right-hand side of the data. Figure 7 shows that on the range from zero to 270 hours, the lognormal distribution has a good fit; however, the data points upwards from 270 hours do not fit well. The histogram of Table 8 shows that the data between 270 and 360 are uniformly distributed. For this reason, we split the emergency lead time distribution between the first part and a second part, as can be seen in Table 1. Now, with probability p = 33/34, the emergency time is lognormal distributed and with 1 – p, the emergency lead time is higher than 270 hours. We determine this probability by using historical data from the end-to-end database.

Evaluate the probability distributions

To evaluate the fitted distributions, we use a Chi-squared test to determine the goodness of fit. This is the most common test to investigate if a certain distribution has a good fit, see for example Buijs (2008). According to this test, the actual data follows the fitted distributions and have a p-value below 0.01, which indicates a good fit of the tested distribution. Employee 6, 7 and 10 validate this information.

Table 2 Fitting probability distributions

Lead- time	Mean (hrs)	Variance (hrs)	Best fitted distribution	Parameters
Local	1.04	0.73	Lognormal	$\mu = -0.27, \sigma = 0.81$
Lateral	16.50	147.22	Lognormal	$\mu = 2.54, \sigma = 0.74$
Emergency [0 – 270)	52.72	7791.99	Lognormal	$\mu = 3.30, \sigma = 1.09$
Emergency (270-360)	314.78	749.29	Continous uniform	$a = 270,\ b = 360$

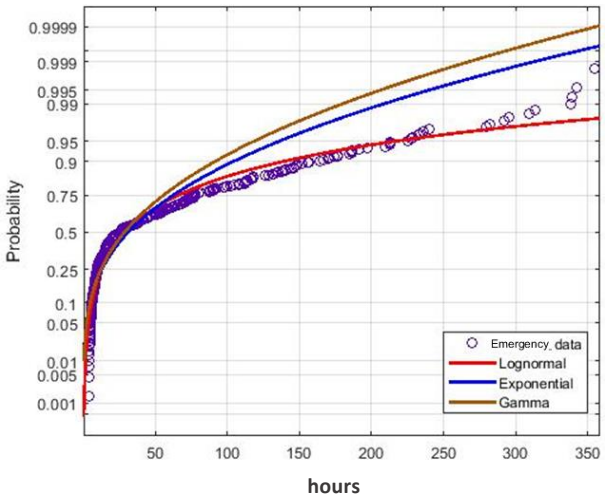


Figure 7 Probability plot: lognormal, exponential and gamma distribution fit on emergency lead time

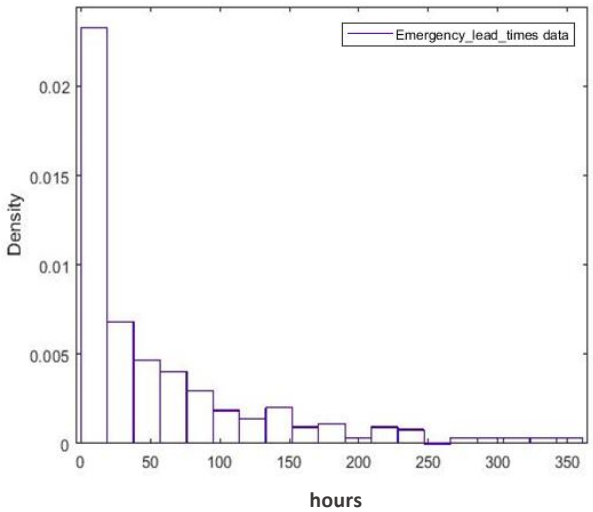


Figure 8 Histogram of emergency time distribution

4.3 Spare parts service measures case study

In this section, we develop a case study to analyse the model as described in Section 4.1. As explained, this model from literature assumes exponentially distributed lead times to determine the optimal base-stock levels. However, as described in Section 4.2, lognormal distributed lead times have a better fit in reality. By applying lognormal distributed lead times, an Erlang loss system (Equation 4.9) does not lead to optimal base-stock levels. However, in literature, there are no methods described to calculate the optimal base-stock levels. Therefore, we first calculate the optimal base stock levels based on exponentially distributed lead times with SPartAn. Then, we analyse the spare part supply chain by simulation and discuss the impact of lognormally instead of exponentially distributed lead times. So, the modeling technique we apply is simulation, which Robinson (2014, p. 4) defines as:

"Experimentation with a simplified imitation (on a computer) of an operations system as it progresses through time, for the purpose of better understanding and/or improving that system."

Besides using SPartAn to calculate the base stock levels, this tool is also used to validate the simulation output. In this section, we first provide the assumptions and the input variables of the simulation model in Section 4.3.1. Then, we discuss the performance and the validation of the simulation model in Section 4.3.2. Finally, we provide the general findings and the main findings of the simulation model in Section 4.3.3 and 4.3.4, respectively.

4.3.1 Input variables of the simulation model

In this section, we first discuss the input variables, based on the supply network of ASML, in Table 3. Then, we provide the different cost input values. All assumptions in this section are validated by Employees 2, 3 and 12 of the service management department.

Table 3 Input data

Variable	Description
The local (t^{loc}), lateral (t^{lat}) and emergency (t^{em}) lead times.	For these three lead times, we assume the probability distributions as described in Section 4.1.
Replenishment lead time t^{rep}	For every spare part at ASML, there is a different expected replenishment time, which is regardless of the distance of a warehouse. The average time of the replenishment lead time is 336 hours (Employee 5). Furthermore, we use the exponential distribution to model the replenishment lead time since this is common in literature, see e.g. Van Houtum & Kranenburg (2015).
Amount of machines at a plan group (Z_n)	We assume 15 machines for plan group n and the same number of machines for the plan group of the main warehouse. This is the most common number of machines for a plan group.
Mean waiting time for a spare part agreement for the main warehouse	If the main warehouse has a mean waiting time for a spare part agreement, we assume this is a 1% target rate since this is the most common target
Fill rate agreement for the main warehouse	If the main warehouse has a fill rate agreement, we assume this is a 95% target rate since this is the most common service level fill rate.

We calculate the stock levels of the warehouses with SPartAn where we use the run of September 2018. Furthermore, we assume that the demand for spare parts (the failure rate of a part) follows a Poisson process, which is common in the literature, see e.g. Basten & Van Houtum (2014). Moreover, we do not make a difference between older and newer machines based on failure rates. We calculate the failure rates per SKU based on the historical data of the total install base of NXT machines of the last 9 years. As mentioned in Section 4.1.2, multiple costs should be considered within the simulation model. These

variables with the corresponding values and description are graphically shown in Table 4. We normalize the costs of the transshipments, where the emergency transshipment cost is set at 100.

Table 4 Cost input variables

Variable	Value	Description
Holding cost rate R	17 %	This percentage is assumed by ASML regardless of the stock location. The holding cost rate is the sum of 8% weighted average cost of capital and 9% of overhead, holding, and other costs (Employee 4).
Cost of emergency transshipment C^{em}	100	Within ASML, it is assumed that an emergency transshipment costs 100% per delivery since this is the average emergency cost to send an order.
Cost of lateral transshipment C^{lat}	42	Within ASML, it is assumed that a lateral transshipment costs 42% per delivery since this is the average lateral cost to send an order.
Cost of replenishment transshipment C^{rep}	0	There are no costs considered for the replenishment transshipments since the number of these deliveries is the same for every scenario.
Cost of local transshipments C^{loc}	0	There are no costs considered for the local transshipments since the number of these deliveries is the same for every scenario.

4.3.2 Running and validation of the simulation model

The simulation model is a discrete-event-based simulation and has been developed using the software MATLAB, whereby the single model runs within fifteen minutes. In addition, the multi-inventory model runs within forty minutes using an Intel Core i5 processor and a parallel pool add-on. In order to provide a well-performing simulation model with a reliable output, we first describe how the warm-up time is determined. Then we explain the procedure to determine the number of replications and finally, we show how we evaluate the simulation model.

Warmup time

When the simulation starts, no SKUs are in transit and the inventory stock levels are the same as the base stock levels. However, this is not the steady state behavior of the model, which leads to estimation in the output of the simulation. In literature, this is known as the problem of the initial transient or the start-up problem (Law, 2007). In order to minimize this problem, we determine a warm-up period whereby we follow Law (2007). This author discusses the following steps:

1. Make n replications of the simulation (where n must be greater or equal to 5) and with a relatively large length m. This leads to output Y_{kq} with k is the k^{th} observation from the q^{th} replication.
2. Create an average process with,

$$\bar{Y}_k = \sum_{q=1}^{n} Y_{qk}/n \qquad (4.14)$$

3. Smooth the average process output ($\bar{Y}_1(w), \bar{Y}_2(w), \ldots. \bar{Y}_{m-w}$) from all replications with a moving average whereby w is the number of time points on which an average is taken. For the formulas to determine w, we refer to Law (2007, p.510). In summary, $\bar{Y}_k(w)$ is the average of 2w + 1 observations of the averaged process output. This method is called the moving average since i is moving through time.
4. Plot the moving average and we choose the warmup time beyond the time that the moving average appears to converge. It is better to choose the warmup time too long rather than too small.

We evaluate the stock points and analyse when these show a steady-state behavior, See Appendix E. In other words, we determine the warm-up period after the point that the moving average is converged. We conclude that the warm-up time is set for 30 weeks and we run the model for the next 10 years

Number of replications

To determine the number of replications, also known as sub-runs, we follow Byrne (2013). This author state that the number of replications in a simulation model is given by a fixed sample size procedure. The goal of this procedure is to have the minimum number of replications whereby one percent (w = 0.01) of the average of the replications falls within a 95% confidence interval ($z_{\alpha/2}$ value of 1.96). Since we assume a normal distribution, we need as minimum 30 replications. We first determine the average of the sample group with the formula:

$$\bar{\mu} = \frac{\sum_{q=1}^{Q} \mu_q}{n} \tag{4.15}$$

And the point variation with,

$$S^2 = \frac{\sum_{q=1}^{n} (\mu_q -)^2}{n-1} \tag{4.16}$$

Then we calculate the coefficient of variation with the following formula:

$$CV = \frac{S}{\mu} \tag{4.17}$$

Finally, we can determine the minimum number of needed replications with,

$$Q = \left(\frac{z_{\alpha/2}}{w} CV\right)^2 \tag{4.18}$$

By following these formulas, we conclude that a minimal number of replications is 101. The calculations are given in Appendix F. With 101 replications, 95% of the outcomes of the different sub-runs is in between 1% below and 1% above the mean value.

Validation of the simulation model

We validate the simulation based on the article of Sargent (2011). This author discusses three applicable methods for our research to validate the simulation model. These are:

1. Comparison to other models
2. Extreme conditions test
3. Face validity

By using the first method, we validate the service measures mean waiting time for a spare part and fill rate with SPartAn. Appendix G shows that, when we apply exponential lead times in the simulation model, the maximum deviation is 1%. This is due to one situation in the simulation model, which is different in SPartAn. Namely, when we do not have SKU i on stock at a local warehouse when a demand of this SKU arrives, we sent an emergency transhipment. However, if the replenishment order in transit arrives earlier than the emergency transhipment, the replenishment order is delivered and the emergency order is placed on stock. This leads to lower mean waiting time for a spare part level of our simulation model in comparison with the mean waiting time for a spare part values of SPartAn. However, this difference is

minimal. Furthermore, we test the model based on the extreme condition test, were we input high base stock levels and long processing times to analyse the service measures in Appendix H. The face validity is done by analysing the waiting times. By performing this method, no notable results were found. In other words, the simulation model works as it should.

4.3.3 General findings of the simulation model

In this section, we discuss the general findings of the simulation model. We start by explaining the influence of the lead times (including emergency, lateral and local lead times) on the SLA of the main warehouse. In the model only scenario 2a (which has a main warehouse with a mean waiting time agreement) and 3a (the main warehouse has a fill rate agreement) are influenced by the main warehouse, therefore we compare these two models. This comparison is done by the mean waiting time for a spare part and the corresponding fill rate. The mean waiting time for a spare part is given in the percentage of the total waiting time for a spare part from the total available time. So, the lower the percentage the better for the customer. A lower the corresponding fill rate the better for ASML since less stock is required, which leads to lower holding costs. In order to analyse the impact on the complete network, we sum up the individual service measures for both warehouses (local and main) and divided these by two. This results in an overview of the average service measures over both the warehouses for scenario 2a and 3a. Finally, we discuss the impact of the machines at a plan group. We provide the findings underlined and as the title of each paragraph, in which we further explain this specific finding.

<u>Based on different emergency lead times, scenario 2a is more beneficial for ASML and its customers,</u>
Table 5 shows the mean waiting time for a spare part and the corresponding fill rate level of Scenario 2a and 3a by different emergency lead times. We conclude that scenario 2a is more beneficial based on a low fill rate level and a low mean waiting time at the emergency times 24 and 48 hours. By applying an average emergency lead time of 72 hours, the required fill rate is higher. However, in practice, this is never applied by ASML. Also, the mean waiting time is more than three times higher, while the corresponding fill rate level is only 0.2 percent lower for scenario 3a. Therefore, we recommend a mean waiting for a spare part agreement for the main warehouse (scenario 2a) regardless of the emergency time.

Table 5 Influence emergency time

Emergency lead time				
	Scenario 2a	Scenario 2a	Scenario 3a	Scenario 3a
Average emergency time local (hrs)	Percentage of the mean waiting time	Corresponding fill rate level	Percentage of the mean waiting time	Corresponding fill rate level
24	1.0%	91.6%	3.2%	92.9%
48	1.0%	94.6%	3.2%	94.7%
72	1.0%	95.4%	3.2%	95.2%

The lateral transshipment time has a negligible effect

Table 6 shows the impact of the lateral lead time on scenario 2a and 3a. As can be seen, different lateral lead times have a small influence on the scenarios. This is due to the fact that in both scenarios, almost the same number of lateral transhipments occur.

Table 6 Influence lateral lead time

Lateral lead time				
	Scenario 2a	Scenario 2a	Scenario 3a	Scenario 3a
Average lateral time to local (hrs)	Percentage of the mean waiting time	Corresponding fill rate level	Percentage of the mean waiting time	Corresponding fill rate level
10	1.0%	93.6%	3.2%	93.6%
15	1.0%	94.5%	3.2%	94.4%
20	1.0%	95.0%	3.2%	95.1%

Based on different local lead times, scenario 2a is more beneficial for ASML and its customers.

Table 7 shows the results of the influence of different local lead times on scenario 2a and 3a. We note that scenario 2a has both a lower corresponding fill rate and a lower average mean waiting time when the local lead times are 0.5 and 1 hour. In the case that 1.5 hours is applied, scenario 3a has a lower corresponding fill rate. However, this scenario has a higher average mean waiting time in comparison to Scenario 2a. Therefore, scenario 2a is more beneficial for ASML and its customers.

Table 7 Influence local lead time

Local lead times				
	Scenario 2a	Scenario 2a	Scenario 3a	Scenario 3a
Average local time both warehouses (hrs)	Percentage of the mean waiting time	Corresponding fill rate level	Percentage of the mean waiting time	Corresponding fill rate level
0.5	1.0%	93.5%	3.1%	93.9%
1.0	1.0%	94.6%	3.2%	95.0%
1.5	1.0%	95.8%	3.3%	95.5%

The number of machines has a negligible impact on the required fill rate level.

The more machines the higher the fill rate level that is required to obtain a 1% mean waiting time for a spare part level since there is a higher probability that a slow-moving part will break (these parts are often not stocked locally). This results in more variety of SKUs on stock and thus higher inventory costs per machine. However, this difference turns out as negligible. The analysis is given in Appendix I, where we compare the required fill rate levels by 5, 15 and 50 machines.

4.3.4 Main findings of the simulation model

We start this section by providing the corresponding values of the different scenarios as described in Section 4.1.3. These corresponding values are required to, in an equivalent way, compare the different scenarios. Next, we discuss the waiting time on spare part probability distributions, which results from simulating the different scenarios. Finally, we use the waiting on a spare part time probability distributions to explain the time to repair distribution and provide the coefficient of variation of these distributions. Figure 9 shows an overview of the scenarios.

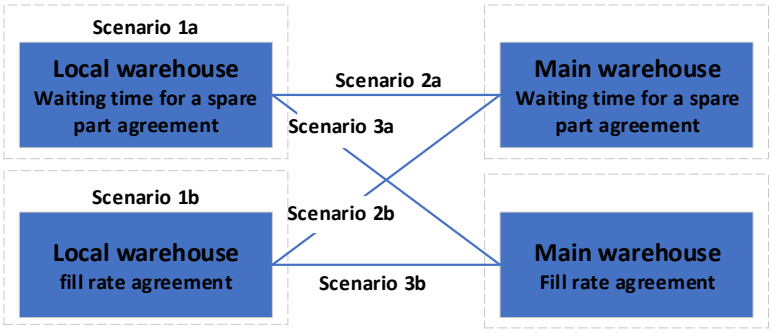


Figure 9 The investigated scenarios

Corresponding values for the different scenarios

We calculate the corresponding values of the different scenarios using SPartAn. First, we provide the 'a' scenarios that are optimized based on a 1% mean waiting time for a spare part agreement at the local warehouse. Then, we discuss the 'b' scenarios, which we analyse based on a fill rate agreement. The 'b' scenarios have the same fill rate level. However, since the emergency time is longer the corresponding mean waiting time for a spare part of the local warehouse is higher.

This leads to the results as shown in Table 8. Note that adding the main warehouse has reduced the required fill rate level whereby the performance (mean waiting time for a spare part level) remains the same. This is due to the fact that some orders can be delivered by a lateral transshipment instead of an emergency transshipment. Therefore, the local warehouse is able to hold less stock.

Table 8 Corresponding fill rate levels to 1% mean waiting time for a spare part level

Scenario	Fill rate level local	Mean waiting time level for a spare part local	Supply chain	SLA local	SLA Main	Assumed emergency time t^{em}
1a	96%	1%	Single	Mean waiting time	N.A.	Normal*
2a	94%	1%	Network	Mean waiting time	Mean waiting time	Normal*
3a	95%	1%	Network	Mean waiting time	Fill rate	Normal*
1b	96%	4.2%	Single	Fill rate	N.A.	336 hrs.
2b	94%	14.3%	Network	Fill rate	Mean waiting time	336 hrs.
3b	95%	11.4%	Network	Fill rate	Fill rate	336 hrs.

*We denote normal as the probability distributions as given in Section 4.2.

Moreover, we conclude that scenario 2a leads to the lowest required fill rate level, based on the investigated supply chain model. This is due to the fact that the main warehouse with a mean waiting time for a spare part agreement holds more inventory and thus is able to send more lateral transshipments. However, there is a tipping point based on the different lead times, which results in scenario 3a as the optimum scenario. This tipping point occurs when the emergency lead time is below 40 hours and the local and lateral transshipment times remain the same. Furthermore, we conclude that the 'b' scenarios have higher corresponding mean waiting time for a spare part values since we optimize these scenarios based on a local fill rate service measure. The mean waiting time for a spare part level increases dramatically when fill rate levels decrease.

Cost of the different Scenarios

The costs are calculated as described in Section 4.1.2. In order to compare the costs of the single location (scenario 1a and 1b) with the network models, we double the holding costs (this is shown in the column main holding costs of Table 9). This is reasonable since, in the other scenarios, the main warehouse has the same spare part demand from its plan group. Table 9 shows all normalized costs per component for the different scenarios. The total cost of scenario 1a is set as 100.

The difference in costs between the corresponding scenarios (such as 1a and 1b) are proportionately smaller than the costs of the different service measures (such as 2b and 3b). This is due to the fact that transportation costs are a relatively small fraction of the total costs in comparison with the holding costs. This is, among other things, pointed out by Van Houtum & Kranenburg (2015). We conclude that scenario 2b has the lowest costs followed by scenario 2a. These two scenarios have the lowest costs because they have the most stock at the main warehouse, which is designated for both plan groups. Because of the higher stock levels at the main warehouse, the risk of out of stock is reduced. Therefore, less stock in total is required.

Table 9 Costs per scenario normalized

		Local holding costs	Main holding costs	Lateral transshipment costs	Emergency shipment costs	Total costs
Scenario	1a	48.5	48.5	0.0	3.0	100.0
	1b	48.5	48.5	0.0	0.0	97.0
	2a	34.4	50.3	0.6	0.6	85.9
	2b	34.6	48.5	0.0	0.0	83.1
	3a	40.0	45.8	0.3	0.9	87.0
	3b	38.4	48.5	0.0	0.0	86.9

Waiting on spare part time probability distribution

The different scenarios lead to corresponding waiting on spare part probability distributions, which are graphically displayed in Appendix J. In addition, the different scenarios lead to a different number of extreme long downs (XLDs) as well. As previously described in Section 3.4, we are not investigating the XLDs service measures. However, we are able to provide the number of XLDs per scenario to see the impact of the different service measures.

Table 10, on the one hand, shows the normalized results of the simulation whereby we set the mean of 1a to 1%. On the other hand, the number of XLDs are given. We first note that all 'a' variants have a lower variation than their corresponding 'b' variant. This makes sense since the variants have a lower emergency lead time. In addition, we note that 2a has the lowest standard deviation of all scenarios. This is the result of the high number of lateral transshipments that is applied in comparison with the other scenarios. From all scenarios, 2b has clearly the highest standard deviation. This makes sense since this scenario has the lowest fill rate level and the main warehouse will not send lateral transshipments.

Table 10 Waiting on spare part time distribution parameters

Scenario	1a	1b	2a	2b	3a	3b
Mean	1%	4%	1%	14%	1%	11%
std. Def.	6%	25%	5%	79%	6%	62%
Number of XLDs per machine per year	100%	138%	163%	250%	150%	225%

Based on the number of XLDs, we conclude that downtimes longer than 12 hours occur less at single stock scenarios (1a and 1b). This is due to the fact that, on average, a lateral transshipment takes 16.5 hours and thus often leads to an XLD. Especially, since the fill rate level of the local warehouse is lower at the multi-stock level scenarios instead of the single stock scenarios.

The coefficient of variation of the time to repair

We express the variation of the system availability in the time to repair the coefficient of variation (c_r), as noted in Section 3.1. The c_r is more meaningful in relationship to the standard deviation since it is the relative variability and, therefore, multiple scenarios can be compared. Furthermore, the time to repair exists of multiple elements (as noted in Section 2.2.2), however, we assume the preparation, maintenance action and the recovery times as deterministic. This means that these parameters are assumed as fixed values and do not have variation. Therefore, we take the mean time of these parameters into account by calculating the c_r, which leads to the following formula:

$$c_r = \frac{\sigma_{waiting\ for\ a\ spare\ part}}{\mu_{preparation} + \mu_{waiting\ for\ a\ part} + \mu_{maintenance\ action} + \mu_{recovery}} \tag{4.19}$$

Table 11 shows the values of the time to repair coefficient per scenario. Hopp & Spearman (2001) state that a coefficient of variation below 0.75 is denoted as low variability, coefficients between 0.75 and 1.33 are set as medium and a coefficient above 1.33 is determined to be high variability. We conclude that the scenarios 1a, 1b, 2a, and 3a classified as low variability and 2b and 3b as high variability.

Table 11 Time to repair coefficient of variation

Scenario	1a	1b	2a	2b	3a	3b
C_r (Value)	0.27	1.00	0.23	2.24	0.26	1.94
C_r (Classification)	Low	Low	Low	High	Low	High

4.4 Conclusions and recommendations of part I

In this section, we first present on the main findings that answer the research question and then provide further conclusions. Finally, we briefly describe the recommendations for ASML based on the first part. The research question of Part I is as follows:

"How do different spare part service measures and their parameters influence the total inventory costs and the probability distribution of the time to repair?"

Table 12 shows the main conclusion of Part I. We conclude that a mean waiting time for a spare part agreement for both the local and main warehouse leads to the lowest coefficient of variation (scenario 2a). This is regardless of different lead times. In addition, this scenario has almost the lowest costs as well. Table 12 shows that scenarios 1a, 1b, 2a, and 3a result in a low coefficient of variation of the time to repair. The other two scenarios, 2b and 3b, lead to a high coefficient of variation. Finally, we conclude that the scenarios with a single stock location have the highest costs.

Table 12 Scenario conclusions

Scenario	**1a**	**1b**	**2a**	**2b**	**3a**	**3b**
C_r (Value)	0.27	1.00	0.23	2.24	0.26	1.94
C_r (Classification)	Low	Low	Low	High	Low	High
Costs	100	97.1	85.4	83.1	87.0	86.9

Furthermore, we conclude:

- The local and lateral lead times fit best a lognormal distribution.
- The emergency lead time fits best a combination of a lognormal distribution for values from zero to 270 hours and a uniform distribution for values from 270 until 360 hours.
- The local lead time has more impact, in comparison with the emergency lead time, on the mean waiting time for a spare part service measure. This is due to the fact that demand always observes a local lead time and only a small part has an added emergency time.
- The number of machines at a planning group has a negligible impact on the difference of service measures. It is true that more stock is placed when a warehouse is responsible for more machines, but, the change in the fill rate and the mean waiting time for a spare part service measures are negligible.
- If the local warehouse has a fill rate agreement, this warehouse can be approached as a single warehouse instead of a part of a network. This is because there is no need to send lateral transshipments to this warehouse since there is no time constraint
- Adding an XLD service measure has an impact on the standard deviation of waiting time in a spare part. Especially on the scenarios where the local warehouse has a fill rate agreement.

Based on all the conclusions we recommend the following:

For ASML and its customers, it is most beneficial to have a mean waiting time for a spare part instead of fill rate agreement. This is due to the fact that this reduces the inventory costs for ASML and leads to a lower coefficient of variation of the time to repair for the customers. Therefore, we recommend that ASML promote the mean waiting time for a spare part agreement. When a customer still wants a fill rate agreement, ASML has to calculate the extra costs and add this in the costs of the SLA contract.

Furthermore, we recommend to consider that single stock locations are more costly than local warehouses that are close to each other or serve multiple plan groups. Single stock locations lead to worse results of performance measures and higher costs since more stock is needed to respond to the risk of unavailabilities. ASML should take this into account when agreeing on a certain mean waiting time for spare parts or fill rate level the and corresponding price of the agreement.

5 The influence of the time to repair probability distribution on the wafer fabrication process

This chapter is the second part of the research and addresses the second research question:

"How do different time to repair probability distributions, resulting from different spare part service measures, influence the work in progress level of a wafer fabrication process?"

In order to answer this question, we start by describing the real-life wafer fabrication process in Section 5.1. Thereafter, in Section 5.2 we provide the mathematical model, which explains the investigated process and the assumptions of the research. In addition, this Section also explains the simulation model where we implemented the assumptions and use the input of the time to repair probability distributions of the different scenarios from Chapter 4. Finally, we discuss the influence of this input on the wafer fabrication process and provide the conclusions in Section 5.3 and 5.4, respectively.

5.1 General information about the wafer fabrication process

Within the four main phases in the semiconductor industry, we investigate the wafer fabrication process because of two reasons. First, Gupta et al. (2006) state that, within the semiconductor manufacturing process, the wafer fabrication is the most costly and time-consuming. Second, ASML's machine is regarded to be the bottleneck of this process.

The wafer fabrication process starts as follows, different layers of material with different electrical characteristics are built on a raw silicon wafer. Each layer consists of transistors that are connected by interconnecting circuits. Moreover, the different layers are built in seven distinctive steps as described by Gkorou et al. (2017) and can be seen in Figure 10. For more details about the fundamentals of semiconductor manufacturing and a more in-depth description of the process, we refer to May & Spanos (2006). To produce a layer, the following process steps are followed (Gkorou et al., 2017):

1. Deposition, a thin film of dielectric material separates the layers in such a way that the different layers are isolated.
2. Photoresist coating, a chemical substance, which is sensitive to light, is spread over the dielectric material.
3. Exposure, light is exposed to a mask which leaves an imprint on the photoresist coating. At the imprint, the photoresist coating evaporates and space arises. The lithography machine, that performs this step, is the most costly and therefore the bottleneck of the process.
4. Developing, the exposed places are further developed and updated.
5. Etching, the work in process layer is immersed in an acid bath in order to remove the dielectric material at the places where the photoresist coating is evaporated. Now the imprint is etched in the dielectric material.
6. Ion implantation, the space in the form of the imprint is filled up with an electrically conductive substance.
7. Stripping, the last step is to remove the residual photoresist coating.

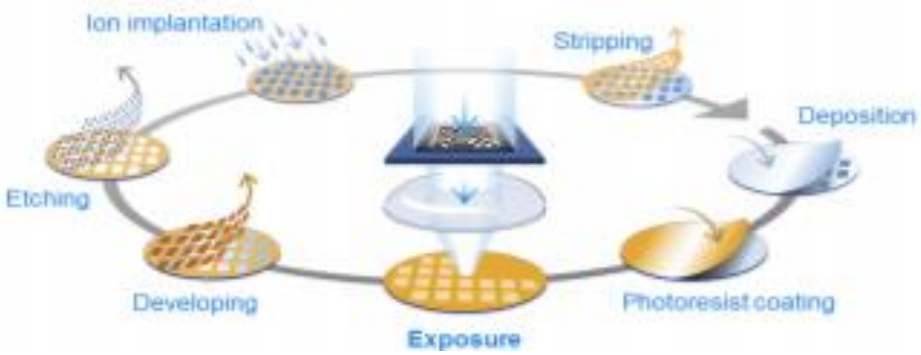


Figure 10 Process steps of a layer (Gkorou, et al., 2017)

Lee et al. (2002) state that a wafer consists of approximately 20-40 layers. Although all layers are different from each other, they follow the same process sequence. So, the same process steps are executed multiple times to produce a wafer. This is different in comparison to other serial manufacturing systems whereby the parts follow a random route (job shops) or a fixed route (flow shops) and does not visit the same machine again. Re-entrant production processes are harder to schedule because of uncertainties in the set-up times, machine failures and fluctuations in the repair time of the machine.

According to Gupta et al. (2006), the process to build the different layers is complex because of (1) the large number of production steps (approximate between 600-800 different steps (Lee, Park, & Kim, 2002)), (2) re-entrant (recirculating flows), (3) random equipment failures, (4) sequence-dependent tool setups, (5) batch processing tools and (6) expansion of capacity is expensive. In a study about the improvements of the cycle-time in this process, Akcalt, Nemoto, & Uzsoy (2001) describe a scaled-down representation of the production line including production and setup times.

5.2 Wafer fabrication model description

In this section, we explain the wafer fabrication model, which is a scaled-down representation of the process as described in Section 5.1. First, we provide the general model and the performance measures in Section 5.2.1. Then, in Section 5.2.2, we explain the different process steps within the wafer fabrication process and finally, we summarize the most important assumptions of the model in Section 5.2.3.

5.2.1 Wafer fabrication process

Since the wafer fabrication process, as described in Section 5.1, is very complex and difficult to model, researchers investigate scaled-down representations of the total process. Hung & Leachman (1999) starts to only model the bottleneck processes in detail and model the rest of the processes more globally (as delay). Many authors followed this approach because a really detailed focus on the bottleneck processes leads to high representative results (Monch, Uzsoy, & Fowler, 2017a). Figure 11 shows our developed model, which is based on the research of Akcalt et al. (2001) and Lin & Lee (2001).

The goal of the model is to investigate the following performance measures based on different repair time probability distributions:

- The average WIP level in the production process
- The average cycle time of a wafer in the process
- The throughput of the process
- The utilization of the bottleneck workstation
- The availability of the machines

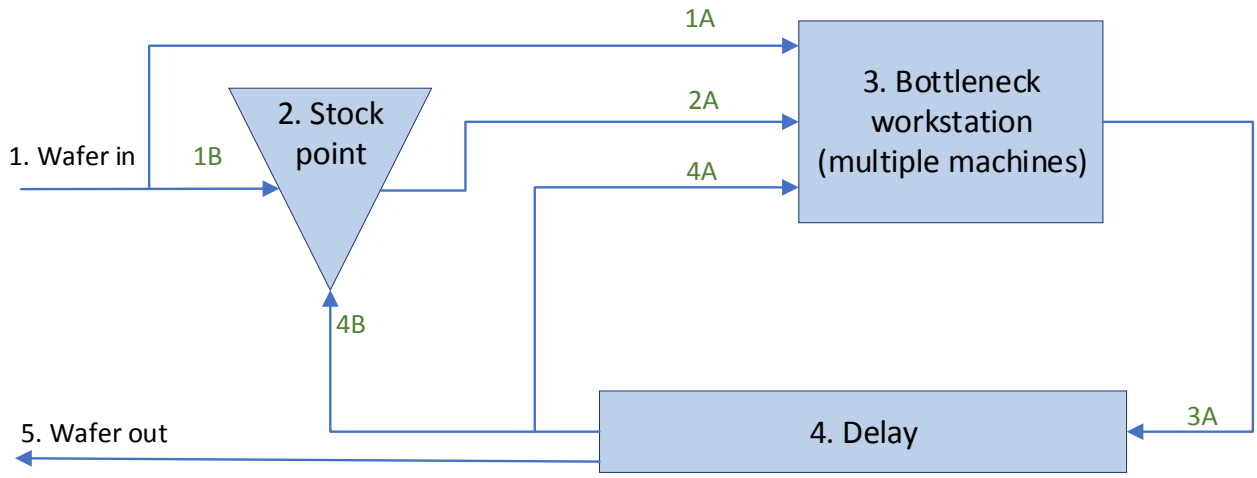


Figure 11 Investigated wafer fabrication process

In general, there are multiple wafer types, which move through the wafer fabrication process in a batch. We assume that each type has the same fixed route. This route is predetermined and consists of multiple workstations. Furthermore, the wafer types differ in the number of arrivals per hour and the number of production rounds. According to Lin and Lee (2001), the number of operations before the bottleneck is relatively low. Therefore, we do not consider these operations and set the bottleneck machine as the starting process of each loop.

5.2.2 Process steps

Within the wafer fabrication process (as shown in Figure 11), the wafers pass the following processes:

1. Wafer arrived at the system

We assume that a batch arrives according to a deterministic process since the inflow of wafers is planned. When a batch arrives at the system this part is processed directly (arrow 1A) or is placed on stock (arrow 1B). In literature, two different dispatch rules to assign a product to a machine are the most commonly used within the wafer fabrication process (Akcalt, Nemoto, & Uzsoy, 2001). On one hand, the flexible assignment policy and on the other hand the dedicated assignment policy.

The flexible policy implies that every batch can be assigned to each machine. Same as in part I, we assume 15 machines. If all 15 machines are occupied, the batch is placed in the stock point (arrow 1B). At this point, the batch is waiting until a machine is available. This assignment policy is easy to work with, however, not a very efficient assignment policy. This is due to the fact that this policy is not taking setup times into account. When a machine is available and just produced type A, it is possible that product type B (whereby a new setup is needed) is assigned to this machine. However, it is possible that the next available machine just has produced type B. In this case, it is better not to assign product B to the first

machine and wait until the second machine is available. To overcome this problem, Akcalt et al. (2001) describe another dispatching rule, namely the dedicated assignment policy. This policy differs from the flexible assignment since machines are assigned to a certain wafer type. If a batch is ready to enter the workstation, it only checks the assigned machines for this type. In the case that one of these machines is available, the batch enters the workstation. Otherwise, the batch is placed in the stock point. When a machine is finished producing a batch, it first checks the stock level of the assigned wafer type. If these are not available, the machine produces another wafer type from the stock point. We assume the dedicated assignment policy since this is a more efficient dispatching rule and more commonly used in practice.

2. Stock point

We assume that the stock point follows a First Come First Serve policy with preemption and has infinite stocking places. In the case that a bottleneck machine is available, it checks the dedicated wafer type, the batch of this corresponding type that arrives first is sent to the bottleneck workstation (arrow 2A). If the machine is idle and the dedicated wafer type for this machine is not on stock, it processes the batch that arrives first.

3. Bottleneck station

The bottleneck station consists of multiple NXT machines. In our research we assume 15 since this the most common number of machines (Employee 2 and 3). Furthermore, we assume the processing time as deterministic. Based on the failure times of the machines within the bottleneck process, we use the time to failure data from part 1. In addition, we assume that only producing machines can break down. When the machine breaks down, there is a repair process. We assume that the maximum number of machines that can be repaired at the same time is fixed. The repair time is based on two factors; firstly, the deterministic components (preparation, maintenance action, and recovery) as noted in Section 2.2.3 and secondly, the waiting time probability distribution. This repair lead time is the result of part I of this research. The machine starts producing once it is repaired. When a wafer is finished at the bottleneck machine, it is sent to the delay process. Besides, the machine verifies if there is a batch in the stock point that can be produced. If this is the case, the machine continues to produce, otherwise, the machine is idle.

4. Delay process

This process consists of all non-bottleneck processes including stripping, deposition, placing the photoresist coating, developing, etching and ion implantation, as graphically displayed in Figure 12. We follow Akcalt, Nemoto, & Uzsoy, (2001) by modelling these stations assuming deterministic processing times and parallel identical servers with a buffer of infinite size.

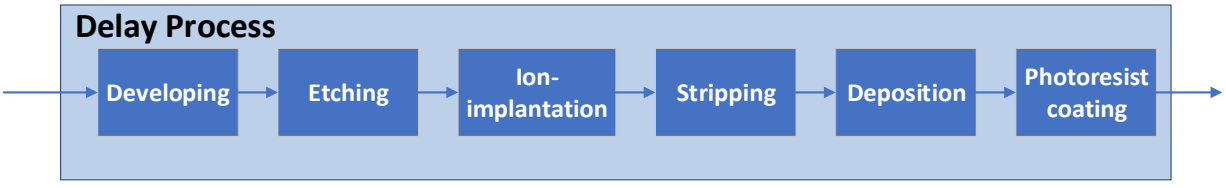


Figure 12 Delay process

When a wafer finishes the delay process, there are three different options:

- The wafer has completed all the production loops and it leaves the system,
- The wafer is not finished and there is an available dedicated machine for this specific wafer type. Then the wafer is transferred to this station (arrow 4A).
- The wafer is not complete and there is no designated machine available. The wafer is transferred to the stock point (arrow 4B).

5. Wafer leave the system

The wafer leaves the system if it completes all the production loops. This is the last step of our modeled process.

To analyse the performance measures, it is important to have a balanced production process. This means that the number of wafers that arrive in the system is less than the maximum output. In this case the number of WIP components within the system increase to infinity over time. We develop the following formula to investigate if the line is balanced:

$$\sum_{b}^{B} \gamma_b + \sum_{b}^{B} \gamma_b * l_b < (Z * A) * \sum_{b}^{B} r_b \qquad (5.1)$$

Where we denote the following variables:

- b as wafer type $b \in B$
- γ_b as the arrival intensity of the wafer type $b \in B$
- l_b as the number of loops of wafer type $b \in B$
- A as the average availability of the bottleneck process
- r_b as the average production time per batch of wafer type $b \in B$
- Z as the number of machines within the bottleneck workstation

Finally, the production time per batch consists of the production time (t_b^{prod}), probability on a setup (p_b^{setup}), and the setup time (t_b^{setup}) for wafer type b. This is displayed with the formula:

$$r_b = t_b^{\text{prod}} + \sum_{b}^{B} p^{\text{setup}} * t_b^{\text{setup}} \qquad (5.2)$$

A setup only appears when a machine changes the sort of wafer. Otherwise, there is no setup time, which is often the case.

5.2.3 Summary of the assumptions about the wafer fabrication model

In this section we briefly summarize the most important assumptions, which are:

- The batch arrival time and processing times of both the bottleneck process and non- bottleneck processes are all assumed as deterministic.
- We assume a dedicated assignment policy.
- Only the component waiting time for a spare part is stochastic; the other components of the time to repair probability distribution are assumed as deterministic.
- We do not consider transportation times between the processes.
- We assume that the number of wafers that arrive at the system together with the wafers that arrive via the loop is smaller than the maximum number of wafers that the bottleneck station can process

5.3 Wafer fabrication case study

In this section, we analyse the wafer fabrication process as described in Section 5.2. We first show that the wafer fabrication process is to complex and extended to model analytically. However, some elements of the process can be modeled, which we provide in Section 5.3.1. Then, in Section 5.3.2, we explain the simulation model and the input variables. Next, we discuss the running and validation of the simulation model in Section 5.3.3. Finally, we discuss the conclusions of the wafer fabrication simulation model in Section 5.3.4.

5.3.1 Analytical modeling of the wafer fabrication process

We analyse the bottleneck process since this is the leading process in the wafer fabrication process. Hereby we start with the repair process, since this is one of the focus points of this research, and then discuss the bottleneck process as a whole. For all queuing theory formulas, we follow Adan & Resing (2015).

Repair process

In order to model this process, we provide three assumptions. The first assumption states that the mean time to failure of the machines occurs with exponential interarrival times. We denote λ as the mean of these interarrival times. This is assumed in the first part as well, however, in Part I we calculate the mean time to failure for every part of the machine individually. By adding the mean values of these distributions, we can create one exponential distribution. The second assumption is based on the service times (repair time). In the first part of the research, we discuss that the repair time follows a lognormal distribution. However, the simulation model shows that an exponential distribution seems closely related. Therefore, we can assume exponential repair times with mean μ in order to have a M/M/c queuing model. Next, the number of parallel identical servers (c) is assumed to be four (Employee 8). This is the number of machines that engineers of ASML can repair at a customer in parallel. We previously assumed that a machine only breaks down if it is processing and not broken yet. However, since these machines form the bottleneck, the aim of the customer is to have a high utilization. Therefore, the stock point before the bottleneck is always filled. So, we assume that a machine is only idle when it is broken.

We model this process as a continuous Markov chain with different state spaces. These state spaces correspond to the number of broken machines (br) $\{0,1, \dots ,15\}$, including all machines that are in the repair process. Figure 13 shows the flow model of this process and the transition matrix is given in Appendix K.

We analyse this system with a local balance equation, this means we calculate the probabilities (p_i) that a process is in phase i by setting input equal to the output for every phase. This results in the following formulas:

$$(15-(i-1))\lambda p_{i-1} = i\mu p_i\,, \quad i = 1, \dots, c, \tag{5.3}$$

$$(15-(i-1))\lambda p_{i-1} = c\mu p_i\,, \quad i = c+1, \dots, 15 \tag{5.4}$$

And since these are all the phases:

$$\sum_{i=1}^{15} p_i = 1 \tag{5.5}$$

Furthermore, we calculate the average number of broken machines by:

$$\sum_{i=0}^{15} br_i * p_i \tag{5.6}$$

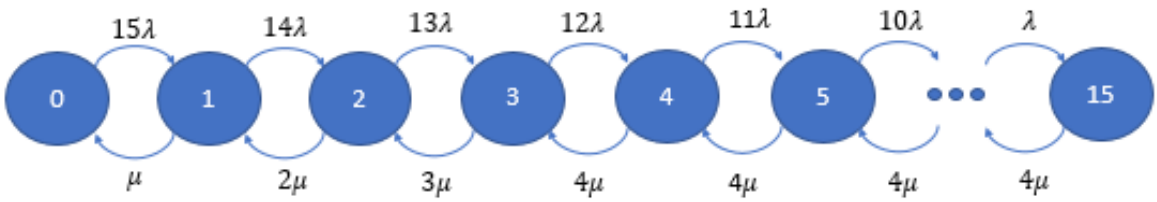


Figure 13 Flow diagram for repair process model

Table 13 shows the outcome of this analysis. Note that, since we assume a steady state and the average waiting time for all 'a' scenarios (mean waiting time for a spare part of 1%). In Appendix L, we show the probability that we are in a certain state for every scenario.

Table 13 Repair process

Scenario	**1a, 2a, 3a**	**1b**	**2b**	**3b**
Probability zero machines are broken (P_0)	5%	4%	2%	2%
Average number of broken machines	3.0	3.5	5.1	4.5
Probability all 4 teams are occupied	34%	44%	71%	62%
Mean availability	80%	76%	67%	70%

Bottleneck process in total

Now we have analysed the repair process of the bottleneck, we are able to analyse the bottleneck process. We first describe the input of the bottleneck process and then the output of this process. The input of the process consists of three streams. Firstly, wafer batches that arrive at the wafer fabrication system. Secondly, if a machine from the bottleneck is available it can take a component from stock if there are parts available. Thirdly, wafer batches that are finishing the delay process and remain into the system .

The output of the bottleneck workstation (see the right-hand side of equation 5.1) is given by :

$$(Z * A) * \sum_{b}^{B} r_b \qquad (5.7)$$

Since the arrival of this process is not following a Poisson distribution and the processing time does not follow an exponential distribution (or other known distribution), we can conclude this is a G/G/c model. Therefore, this model is very complex to solve mathematically (Adan & Resing, 2015). In order to analyse the wafer fabrication system as a whole, we use a simulation model. This is the most common modeling technique in the semiconductor industry (Monch, Uzsoy, & Fowler, 2017a).

5.3.2 The simulation model and the input variables of the wafer fabrication process

In this section, we simulate the model as described in Section 5.1. There are numerous different types of simulations whereby discrete-event simulation (DES) is pointed out as a well-established tool for the wafer fabrication process (Monch, Uzsoy, & Fowler, 2017a). Since detailed simulation models require large amounts of data and long computation times due to the complexity of the process and the number of process steps, several researchers proposed to reduce the level of detail of the simulation model while maintaining sufficiently accurate results. We reduce the computation time by doing an event-based simulation and model the non-bottleneck processes as a delay.

The input values of the simulation model are based on the test bed from the research of Akcalt et al. (2001), which is executed in a wafer fabrication factory in Taiwan. We start by describing the input variables of the characteristics of the wafer types. Then, we provide the processing times and finally, the inflow of the wafers into the system. Each of the wafer types has their own characteristics and input values, which are given in Table 14.

Table 14 Input variables type wafers

Wafer - Type	**A**	**B**	**C**	**D**	**E**
Batch size (units)	25	25	25	25	25
Number of required production rounds (loops)	21	19	20	19	21
Assigned number of machines (number of machines)	4	2	3	2	4
Setup time (hrs)	0.02	0.01	0.02	0.01	0.01

The bottleneck workstation consists of 15 NXT machines, which is the same number as assumed in the first part of the research. Furthermore, we assume a target utilization of 80% and that 4 machines can be repaired in parallel. Based on running the simulation model beforehand we found that the probability on a setup for a wafer is approximate 10%.

We use Akcalt, Nemoto, & Uzsoy (2001) based on the processing times within the wafer fabrication process in order to make our research reproducible. The authors assume that these processing times are deterministic. In recent years these process times are reduced through technological development (the lithography process is executed 12 times faster now (ASML, 2015). However, more up-to-date information is not available in the literature. ASML can provide an average of wafers that a machine can produce, however, a distribution and non-bottleneck of processing times are unknown (Employee 8). This is because customers do not share this information because of competitive reasons. Table 15 shows the mean processing times for each individual process within the wafer fabrication process.

Table 15 Delay mean processing times processes from Akcalt et al. (2001)

Delay process	**Mean process time (Hours per batch)**
Bottleneck process	**1.1**
Delay process	**13.4**
Developing,	4.8
Etching	1.0
Ion-implantation	1.0
Stripping	1.2
Deposition	4.0
Placing the photoresist coating	3.0

So, the maximum capacity of the bottleneck process is the number of batches processed per hour multiplied by the number of machines. This leads to $\left(\frac{1}{1.1}\right) * 15 \approx 13.5$ batches per hour.

Furthermore, the number of arrivals per hour is based on the production rate of the bottleneck workstation, and are shown per scenario in Table 16. As noted in Section 5.2.1, the average wafer inflow in a loop should not be more or equal than the number of wafers the bottleneck is able to produce. Otherwise, the WIP level in the total system will not stabilize over time. We calculate the number of

incoming wafers with the balanced formula as can be seen in equation 5.1 and use the different availabilities as calculated in Section 5.3.1. In order to find an optimum we use a solver as can be seen in Appendix M. It is important to note that the number of loops is of crucial importance to determine the optimal input.

Table 16 Number of batches per hour that arrive at the system per scenario

Wafer type	***Scenario 1a***	***Scenario 1b***	***Scenario 2a***	***Scenario 2b***	***Scenario 3a***	***Scenario 3b***
A	0.095	0.087	0.095	0.079	0.095	0.107
B	0.116	0.108	0.116	0.094	0.116	0.129
C	0.073	0.070	0.073	0.064	0.073	0.087
D	0.116	0.107	0.116	0.095	0.116	0.128
E	0.132	0.119	0.132	0.101	0.132	0.010
Total (batches/hr)	**0.532**	**0.490**	**0.532**	**0.435**	**0.532**	**0.461**

5.3.3 Running and validation of the wafer fabrication simulation model

The simulation model has been developed by using the software MATLAB, whereby a single run takes approximately 1 hour. This is due to the high number of wafers that is inside the process. As hardware, we used an Intel Core i5 processor. Furthermore, the simulation can be described as an event-based simulation model. We first describe the warm-up time of the model. Next, we explain the number of replications and the validation of the simulation model.

Warm-up time

The model is started empty and idle and form the initial stage. In order to eliminate the initial bias, we discard the first 500 hours of the run. We determine initial transient (500 hours) by visual inspection of the output data of the simulation model. Furthermore, we use the moving average and the method as described by Law (2007). We previously described this method for the first simulation model in Section 4.3.2. We run the model for 2 years and we set the warm-up period at 500 hours. As can be seen in Appendix E, this is the moment that the number of wafers is stabilized.

Number of replications

To determine the number of replications, we use the same formula's as noted in Section 4.3.2. We conclude that 46 is the minimum number of replications required as shown in Appendix F. Furthermore, we set w as 0.1 and $z_{\alpha/2}$ as 1.96.

Validation of the model

To validate the model, we again use the methods as described by Sargent (2011). However, since there is no comparable simulation model we only use the extreme value and face validity methods. We took the following steps:

- If we set maximum values for the input values, the results are as expected. For example, if we set the inflow of wafers into the system extreme high (5 times normal the normal value), this leads to an increase of wafers in the stock point, high utilization and an increasing WIP level, as shown in Appendix H.
- We test the Availability results from the analytical modelling (Section 5.4) against the availability from the simulation. These results are closely related.

5.3.4 Findings of the wafer fabrication simulation model

The customers of ASML strive for the best results on their performance measures, which means a high throughput of wafers low WIP level and short cycle time. Table 17 shows the results of the simulation model per scenario. In order to have a balanced line, the input for the 'b' scenarios is different (See Table 16). Therefore, the WIP and cycle time are not good indicators to determine the performance of the 'b' scenarios. However, the difference In throughput is notable. As can be seen, the throughput of the 'a' scenarios is higher in comparison with the 'b' scenarios.

Table 17 Main results simulation Part II

Scenario	Throughput (batches)	WIP level (batches)	Cycle time (hrs)	Availability	Utilization
1a	0.53	38.53	147.72	80%	78%
1b	0.49	29.62	58.57	77%	74%
2a	0.53	27.83	52.18	81%	77%
2b	0.43	92.06	215.19	67%	63%
3a	0.53	34.66	65.02	81%	78%
3b	0.46	122.35	265.59	68%	67%

It is important to note that the throughput of the system is not the same as the throughput of the bottleneck station. This is due to the loop in the system whereby batches visit the bottleneck station multiple times. On average, a wafer visits the bottleneck 20 times (number of loops) and therefore the throughput of the system is approximate 1/20 of the bottleneck capacity. The throughput is the same as the sum of the wafers that arrive at the system per hour (Table 16) since we obtain a balanced line.

We show this, as an example, for scenario 1a. As can be seen in Table 17, the throughput of scenario 1a is 0.53 batches per hour. This is equal to the sum of the input of all wafer types per hour ($0.095 + 0.116 + 0.0732 + 0.116 + 0.132 = 0.533$) as shown in Table 16. The throughput of the system is related to the bottleneck throughput as follows; a number of batches per hour processed on the bottleneck station times the availability is equal to the number of batches that flows into the system times the number of rounds. This leads to the following equation: $13.5 * 0.8$ and $20 * 0.53$, which both result in approximately 10.7 batches per hour.

Finally, it is notable that the numerical results about the availability of the bottleneck station, as described in Section 5.3.1, corresponds to the mean availability results of the simulation.

5.4 Conclusions and recommendations of part II

In this section, we discuss the conclusions based on the wafer fabrication model in order to answer the research question of Part II, which is formulated as follows:

"How do different time to repair probability distributions, resulting from different spare part service measures, influence the performance measures of a wafer fabrication process?"

Table 18 shows the results of the simulation model, where we rank how well the scenarios score based on the performance measures with 1 being best and 6 being the worst. We provide this rank by first taking the scenarios with the highest throughput. Thereafter we rank them based on the lowest WIP and shortest cycle time. We conclude that scenario 2a, local warehouse and main warehouse with both an mean waiting time for a spare part agreement, leads to the best performance measures and the lowest coefficient of variation of the time to repair.

Furthermore, All 'a' scenarios have almost the same coefficient of variation and, therefore, the same wafer input (see Table 16). We conclude that, which is also pointed out by Hopp & Spearman (2001), the performance measures deteriorate by a higher coefficient of variation of the time to repair.

Table 18 Results of the simulation model

Scenario	Performance measures rank	CV time to repair
1a	3	0.27
1b	4	1.00
2a	1	0.23
2b	6	2.24
3a	2	0.26
3b	5	1.94

Based on the conclusions, we recommend to ASML to inform the customer how different service measures influence the CV of the time to repair and thus their process. In this way, there is more understanding of the service measures that a customer can choose and ASML can better serve its customers. Also, ASML is able to reduce costs since a mean waiting time for a part leads often to less stock in comparison with a fill rate agreement. In summary, by informing the customer about the impact of the different contracts, ASML achieve a higher customer satisfaction and reduce costs.

6 Conclusions and recommendations

In this chapter, we provide the results and conclusion of the research in order to answer the following research question:

"How do different spare part service measures and their parameters influence the performance measures of a wafer fabrication process?"

Firstly, we provide the conclusions in Section 6.1. Then, in Section 6.2, we discuss the main limitations of this research. Next, we describe the recommendations on theoretical and practical ground and the options for further research in Section 6.3. Finally, in Section 6.4, we discuss the academic contribution of this research.

6.1 Conclusions

We first describe the main conclusion of this research which provides an answer to the main research question. Then we conclude the remaining results of Part I followed by the results of Part II. More specific conclusions for Part I and Part II are given in Sections 4.4 and 5.4, respectively. For the overview, we again show the different scenarios we researched in Figure 14.

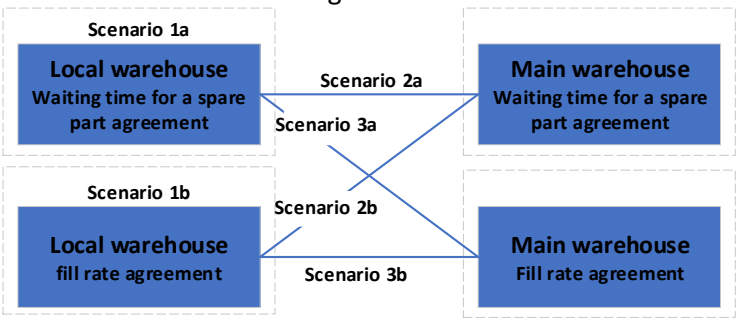


Figure 14 Researched scenarios

Main conclusion and answer to the research question:

Table 19 shows the costs and the coefficient of variation of the time to repair from Part I, as well as the system availability and performance measures of Part II. Together, this forms the main conclusion of our research. Based on the performance measures, the most positive score is ranked 1st and the least positive score is ranked 6th. It is notable that a low coefficient of variation leads to a high level of system availability, which leads to a higher rank on the performance measures. In order to answer the main research question, a mean waiting time for a spare part agreement at a local warehouse ('a' scenarios) leads to more positive performance measures instead of a fill rate agreement. The best results on the performance measures are obtained if there is a main warehouse with a mean waiting time for a spare part agreement as well. In addition, this scenario also results in almost the lowest holding costs for ASML. Therefore, we can conclude that by agreeing on a mean waiting time for a spare part agreement for the local and main warehouse (scenario 2a) a win-win situation is reached for both ASML and its customer. This is due to the low coefficient of variation of the time to repair.

Table 19 Main conclusion

	Costs	CV repair	System availability	Rank based on performance measures
Scenario 1a	100	low	79%	3
Scenario 1b	97	low	77%	4
Scenario 2a	85	lowest	81%	1
Scenario 2b	83	high	67%	6
Scenario 3a	87	low	81%	2
Scenario 3b	87	high	69%	5

6.2 Main limitations of the research

We provide the limitations of this research in two parts. First, we describe the main limitations of this research and then discuss the further research limitations of Part I and Part II.

Within the chosen service measures (i.e. mean waiting time on a spare part and fill rate), we do not take into account the extreme long downtime agreement (XLD). This is due to the fact that there is no optimisation method described in the literature to determine the base stock levels for multi-item models based on this agreement. This agreement focusses on long downtimes, which results in a high coefficient of variation of the time to repair. In this research, we conclude that a higher coefficient of variation of the time to repair leads to deterioration of the performance measures. Therefore, this service measure can be more preferred than a waiting time on a spare part agreement.

Furthermore, we assume that an emergency demand of a warehouse with a fill rate agreement is not fulfilled as fast as possible in our study. This is theoretically true based on the agreement that ASML has with its customers. However, in practice, this does not hold since ASML strives for high customer satisfaction and thus delivers a part as fast as possible. This assumption impacts the time waiting on a part and thus also the coefficient of variation of the time to repair and the results of the performance measures. In other words, by changing this assumption, another scenario can result in the most recommended one for ASML and its customers.

Finally, by modeling the wafer fabrication process, we use old fashioned production lead times for the different processes. This is due to the fact that companies are not willing to share this information based on competitive reasons. When it turns out that the lithography machine is less important or not the bottleneck machine, the coefficient of variation of the time to repair has less influence on the performance measures. In this case, the different performance measures of the different scenarios will be less far apart from each other.

6.3 Recommendations and further research

In this section, we provide the recommendations and further research based on the limitations of the research. We first describe the recommendations from a business point of view in Section 6.3.1 and then, in Section 0, we provide the options for further research from an academic point of view.

6.3.1 Corporate recommendations

We first provide the three main recommendations for ASML and then discuss the model extensions in order to develop a more realistic scenario.

The first recommendation to ASML is to provide mean waiting time for a spare part instead of fill rate agreements to its customers since this leads to a lower coefficient of variation of the time to repair for the customer. The coefficient results in more positive results of the performance measures and thus a higher customer satisfaction for ASML. As a matter of course, a customer can choose their own service level agreement within the performance agreement. However, ASML can provide the results of a certain service level agreement and encourage a customer to choose a mean waiting time for a spare part agreement.

The second recommendation for ASML is to keep in mind that single stock locations are more costly than local warehouses that are close to each other or serve multiple plan groups. Single stock locations lead to worse results based of the performance measures and higher costs since more stock is needed to respond to the risk of unavailabilities. ASML should take into account this higher costs by agreeing on certain mean waiting time for a spare part or fill rate level and by arranging the price of a certain service level agreement.

The third recommendation for ASML is to consider the emergency time for a customer with a fill rate agreement. Now, ASML is always sending a spare part as fast as possible to the customer. However, this is not according to the SLA of a fill rate agreement, which does not consist of a time restriction. Therefore, ASML is overperforming for these plan groups.

In order to make the model more realistic, ASML may investigate the following model extensions:

- Investigate the probability distributions of the preparation, maintenance action, and recovery time on the coefficient of variation. These are elements of the time to repair. We only investigate the wait time for a spare part. We expect that if the variation of these elements is high, the coefficient of variation of the waiting time is less important. This means that the results on the performance measures of the different scenarios differences less. ASML can analyse this by data from the maintenance engineers. For now, this data is not efficient stored in a databank. So, the first step is to investigate which data is important and how to store this information.
- Investigate the XLD service measure for multiple parts and their impact on the coefficient of repair time. For now, it is impossible to provide this relationship since there is no method to determine the base stock level for multiple parts. In literature, only three heuristics are given to provide these stocks for a single item and single stock point (Lamghari-Idrissi, Basten, & Van Houtum, 2018). It is possible that this service measure leads to an even lower coefficient of variation of the time to repair and thus to better performance measures at the customer side. In order to investigate the XLD service measure, a heuristic must be developed to determine the required base stock levels. Then, the impact of this service measure on the time to repair probability distribution must be analysed. If the time to repair probability distribution is known, the simulation of Part II can be used to analyse the impact of the XLD service measure on the performance measures.
- We only investigated the results of the spare part service measures based on a steady state. By applying a finite horizon, the base stock levels will be higher in order to get the same spare part service measures. It is interesting to research this impact on the coefficient of variation of the repair lead time. In this case, different time horizons has to be considered as extra variable in the simulation models.

6.3.2 Academic recommendations

This research is based on the idea of Kiesmüller & Zimmerman (2018), who describe the influence of spare parts provisioning on the WIP level of a production process. Since this was the first research based on investigating the relationship between spare parts and a performance measure, this research can be seen as a building block. The authors researched a two machine model, which is a relatively basic process with two machines in serie and one buffer in between. We build further on this research by investigating spare part service measures and extend the production process. Further research should include various model extensions, such as:

- We investigate a wafer fabrication process, however, the influence of spare part service measures can also be researched on other fabrication processes. Since the wafer fabrication process has re-entrant flows, the bottleneck process plays a crucial role. In a normal series fabrication process, the bottleneck is still important however has less impact than in the wafer fabrication process. Therefore, we expect that the results performance measures of the different scenarios are closer to each other.
- We assume only corrective maintenance. However, preventive maintenance, quick fixes, and other maintenance options can also be considered. In this case the coefficient of variation of the time to repair has less impact on the wafer fabrication process. Therefore, we expect that the results of the different scenarios, based on the performance measures, are more closely together in comparison with the founded results from this study.
- We do not describe the relationship of an XLD service measure on the time to repair distribution. This is an option for further research and already noted in Section 6.3.1.

6.4 Academic contribution

In this section, we first describe the academic contribution of this research. In the scientific relevance of this report, Section 3.2, we explain two gaps that this research fills based on our literature review. These are the following:

- The influence of spare part service measures on the time to repair distribution,
- The influence of the time to repair distribution (coefficient of variation) on the performance measures of a wafer fabrication process (production process with recirculating flows).

In the first part of the report, we show that the fill rate and mean waiting time for a spare part service measures result in a different coefficient of variation of the time to repair. To the best of our knowledge, this is not discussed in the literature. We conclude that a mean time waiting for a spare part service measure, instead of a fill rate service measure, always leads to a lower coefficient of variation of the time to repair. In addition, by considering lateral transhipments, this coefficient of variation is even lower.

In the second part, we describe the influence of the coefficient of variation of the time to repair on the performance measures of a wafer fabrication process. Hopp & Spearman (2001) describe the relationship between this coefficient and a production process with machines in series. However, since the wafer fabrication has recirculating flows, the coefficient of variation of the repair time has a different impact on the performance measures of the system. To the best of our knowledge, this relationship is not described. In this research, we show that the statement of Hopp & Spearman (2001) is true for the wafer fabrication process.

References

Adan, I., & Resing, J. (2015). *Queueing Systems.* Eindhoven : Department of Mathematics and Computing Science, Eindhoven University of Technology .

Akcalt, E., Nemoto, K., & Uzsoy, R. (2001). Cycle-Time improvements for photolithography process in semiconductor manufacturing. *IEEE Transactions on semiconductor manufacturing vol. 14*, 48-56.

ASML. (2015, September 29). *Press Releases*. Retrieved December 5, 2018, from ASML Ships New TWINSCAN NXT Immersion Lithography Platform, Delivering Improved Performance for Volume Production at Next-generation Process Nodes; Demonstrated Advancements in Overlay and Focus Uniformity to Address Multiple Patterning Requirements at th: https://www.asml.com/press/en/s5869?rid=52376

ASML. (2018a, June 6). *Customer support*. Retrieved from ASML.com: https://www.asml.com/products/customer-support/maintenance-operations-support/en/s259?rid=6940

ASML. (2018b, March 28). *https://www.asml.com/company/in-a-nutshell/en/s277?rid=51978*. Retrieved from ASML: https://www.asml.com/company/in-a-nutshell/en/s277?rid=51978

Bahai, A. (2017). IEEE International Solid-State Circuits Conference. *Dynamics of Exponentials in Circuits and Systems* (pp. session 1, plenary 1.2). Santa clara : IEEE International .

Banerjea, S. (2017, December 11). *Semiconductor Industry Outlook* . Retrieved from Semiconductor Industry Outlook : https://www.nasdaq.com/article/semiconductor-industry-outlook-december-2017-cm890008

Banks, J., Carson, J., Nelson, B., & Nicol, D. (2001). *Discrete-Event System.* Prentice-Hall: New Jersey.

Basten, R. J., & Van Houtum, G. J. (2014). System-oriented inventory models for spare parts. *Surveys in Operation Research and Management Science 19*, 34-35.

Blanchard, B., Verma, D., & Peterson, E. (1995). *Maintainability: A Key to Effective serviceability and maintenance managment.* London: Wiley-IEEE.

Buijs, A. (2008). *Statistiek om mee te werken.* Groningen: Noordhoff Uitgevers .

Byrne, M. (2013). *How Many Times Should a Stochastic Model Be Run?* Houston : Departments of Psychology and Computer Science.

De Treville, S., Bicer, I., Chavez-Demoulin, V., Hagspiel, V., Schurhoff, N., Tasserit, C., & Wager, S. (2014). Valuing lead time. *Journal of Operations Management* , 337-346.

Dijkman, R. (2017, March 1). *Business Process Management, Process Discovery [powerpoint presentation].* Retrieved from canvas.tue.nl: https://canvas.tue.nl/courses/241/files/folder/lectures?preview=27748

Enders, P. (2004). *Spare parts inventory control in a multi-item, two-echelon network with lateral and emergency shipments, MSc Thesis.* Eindhoven: Eindhoven University of Technology.

Freedman, D., & Diaconis, P. (1981). *On the histogram as a density estimator: L2 theory.* Heidelberg : Springer.

Ge, Q., Peng, H., Van Houtum, G., & Adan, I. (2018). Reliability optimization for series systems under uncertain component failure rates in the design phase. *International Journal of Production Economics, vol 196(C)*, 163-175.

Gkorou, D., Ypma, A., Tsirogiannis, G., Giollo, M., Sonntag, D., & Vinken. (2017). *Towards Big Data Visualization for Monitoring and Diagnostics of High Volume.* Eindhoven: ASML.

Gupta, J. N., Ruiz, R., Fowler, J. W., & Mason, S. J. (2006). Operational planning and control of semiconductor wafer production. *Production Planning & Control*, 639-647.

Hanbali, A., & Van Der Heijden, M. (2013). Interval availability analysis of a two-echelon multi-item system. *European Journal of Operational Research vol. 228*, 494-503.

Hasan, R. (2016, September 22). The voice of Micron. Eindhoven, Brabant, Netherlands.

Hopp, W., & Spearman, M. (2001). *Factory physics.* Boston: McGraw-Hill companies.

Jiang, H. (2012). *Central warehouse planning within the service parts network of ASML, MSc Thesis.* Eindhoven: Eindhoven university of technology.

Johansson, L. (2011). *Mathematical modeling of inventory control systems with lateral transshipments.* Lund : Lunds Universitet .

Kiesmüller, G. P., & Zimmerman, J. (2018). The influence of spare parts provisioning on buffer size in a production system. *IISE Transactions, vol. 50* , 367-380.

Kranenburg, A. A. (2006). *Spare parts inventory control under system availability, PhD thesis.* Eindhoven: Eindhoven university of Technology.

Lamghari-Idrissi, D., Basten, R., & Van Houtum, G. (2018). Spare parts inventory control under a fixed-term contract with a long-down constraint. *Eindhoven University of Technology*.

Law, A. (2007). *Simulation modeling and analysis, fourth edition.* Boston: Mc Graw-Hill Education .

Lee, Y. H., Park, J., & Kim, S. (2002). Experimental study on input and bottleneck scheduling for a semiconductor fabrication line. *IIE transactions, vol 34*, 179-190.

Lin, Y., & Lee, C. (2001). A total standard WIP estimation model for wafer fabrication. *European journal for operation research vol. 131*, 78-94.

Little, J. (1961). A Proof for the Queuing Formula: L = λW. *Operations Research, vol. 3*, 383–387.

May, G. S., & Spanos, C. J. (2006). *Fundamentals of semiconductor manufacturing and process control .* New Jersey: John Wiley & Sons .

MIT.edu. (2015, December 11). *Semiconductor Costs* . Retrieved from Courses: https://ocw.mit.edu/courses/engineering-systems-division/esd-290-special-topics-in-supply-chain-management-spring-2005/tools/SemiconductorCostModelvFinal.xls

Monch, L., Uzsoy, R., & Fowler, J. W. (2017a). A survey of semiconductor supply chain models part I: semiconductor supply chains, strategic network design, and supply chain simulation. *International journal of production research*.

Nahmias, S., & Olsen, T. L. (2015). *Production and operations analysis seventh edition.* Long Grove: Waveland Press .

O'conner, P. D., & Kleyner, A. (2012). *Practical Reliability Engineering, fifth ed.* Chichester: Wiley.

Reijnen, I. (2009). PDeng thesis: Inventory planning for spare parts networks with delivery. *Eindhoven University of Technlogy*.

Sargent. (2011). Verificatyion and Validation of simulation models . *Proceedings of the 2011 Winter Simulation Conference*, 183-198.

Sched, J. (2011). A survey of problems, solution techniques, and future challaenges in scheduling semiconductor manufacturing operations. *Springer Science+Business Media* , 583-599.

Smets, L. P., Van Houtum, G. J., & Langerak, F. (2012). Design for availability: a holistic approach to create value for manufacturers and customers of capital goods. *Syst Sci Syst Eng*, 403-421.

Thompson, G. (1999). *improving maintainability and reliability through design.* London: Professional Engineering publishing.

Van Aken, J., Berend, H., & Van Der Bij, H. (2012). *Problem solving in organizations, a methodolgical handbook for business and management students.* Groningen: Cambridge university press.

Van Aspert, M. (2014). *PDeng: Design of an Integrated Global Warehouseand Field Stock Planning Concept for Spare Parts.* Eindhoven: Eindhoven, University of Science and Technology.

Van Houtum, G. J., & Kranenburg, B. (2015). *Spare Parts Inventory Control under System Availability Constraints.* New York, Heidelberg, Dordrecht, London: Springer.

Appendix A Abbreviations

We provide all abbreviations of this report in Table 20.

Table 20 Abbreviations

Abbreviation	Definition
BPMN	Business Process Management Notation
BPS	Business Problem Solving
CLCS	Closed-Loop Supply Chains
CSD	Customer Service Degree
CT	Cycle Time
CW	Central Warehouse
DES	Discrete-event simulation
DTWP	Downtime Waiting Part
DUV	Deep Ultraviolet systems
EUV	Extreme Ultraviolet systems
Fab	Manufacturer of semiconductors
FCFS	First Come First Serve
FSD	Field Service Defect
KPI	Key Performance Indicators
LW/Loc	Local Warehouse
MW/Main	Main Warehouse
MN	Material Notification
MTBF	Mean Time Between Failure
MTTF	Mean Time To Failure
MTTR	Mean Time To Repair
OEM	Original Equipment Manufacturer
SAP	System, application, and programs. This is the software with all intern data of ASML
SCM	Supply Chain Management
SKU	Stock Keeping Unit
SLA	Service Level Agreement
TU/e	Eindhoven University of Technology
WIP	Work In Progress parts
TH	Throughput
XLD	Extreme Long Downtime, downtime of a machine longer than 12 hours

Appendix B Employee references

Within ASML, we had conversations with the following Employees in order to obtain information or verify the assumptions we made. All names, the department and the sort of contact can be found in Table 21.

Table 21 Employees

Employee	Departement	Contact
Employee 1	CH CS ENG Operational Leader group	Informal meeting
Employee 2	GL SCM CSCM Service Planning	Informal meeting
Employee 3	GL SCM CSCM Service Planning	Informal meeting
Employee 4	GL SCM CSCM Service Planning	Informal meeting
Employee 5	GL SCM CSCM In-Fab	Informal meeting
Employee 6	GL SCM CSCM FMA BL MPS performance	Informal meeting
Employee 7	GL SCM CSCM FMA BL DUV APPS Performance	Informal meeting
Employee 8	CH CS ENG Field Performance Analysis	Informal meeting
Employee 9	GL SCM CSCM Service Planning	Informal meeting
Employee 10	GL SCM CSCM Service Planning	Informal meeting
Employee 11	GL SCM CSCM FMA Do Mat Del Volume	Informal meeting
Employee 12	GL SCM CSCM Service Planning	Informal meeting

Appendix C Optimization problems

In the optimisation models, we set DTWP as the mean waiting time for a spare part service measure and CSD as the fill rate service measure. In order to calculate the stock levels for every service measure, the following optimization problems are drafted:

Optimization problem 1

$$\min \quad C(S) = \sum_{i \in I} C_i(S_i)$$

subject to

$$CSD_n(\boldsymbol{S}) \geq CSD_n^{obj} \quad \forall n \in N$$
$$S_{i,j} \geq S_{i,j}^{start} \quad \forall i \in I, j \in J$$
$$S_{i,j} \in \mathcal{S} \quad \forall i \in I, j \in J$$

Optimisation problem 2

$$\min \quad C(\boldsymbol{S}) = \sum_{i \in I} C_i(S_i)$$

subject to

$$DTWP_n(\boldsymbol{S}) \leq DTWP_n^{obj} \quad \forall n \in N$$
$$S_{i,j} \geq S_{i,j}^{start} \quad \forall i \in I, j \in J$$
$$S_{i,j} \in \mathcal{S} \quad \forall i \in I, j \in J$$

For all optimization models, we designate the set of all possible solutions by $\mathcal{S}$, whereby $\mathcal{S} = \left\{S = \left(S_{i,j}\right)_{i \in I, j \in J} | S_{i,j} \in N_0, \forall i \in I \; and \; j \in J\right\}$. In addition, $S_{i,j}^{start}$ designates the minimal base stock level for SKU i at location j as required by ASML. These are predetermined by the management of ASML. Furthermore, note that the restriction of the fill rate service measure should be equal to or higher than its objective. This is in contrast to the mean waiting time for a spare part, which should be equal to or lower than its objective.

Appendix D Fitting lead time probability distributions

In this research, we plotted a lot of distributions in order to find the best fit including Weibull, normal etc. For the overview, we only show the distributions that we close related to the actual data. We show the local lead time, lateral lead time and the emergency lead time distributions in Figure 15, Figure 16 and Figure 17, respectively.

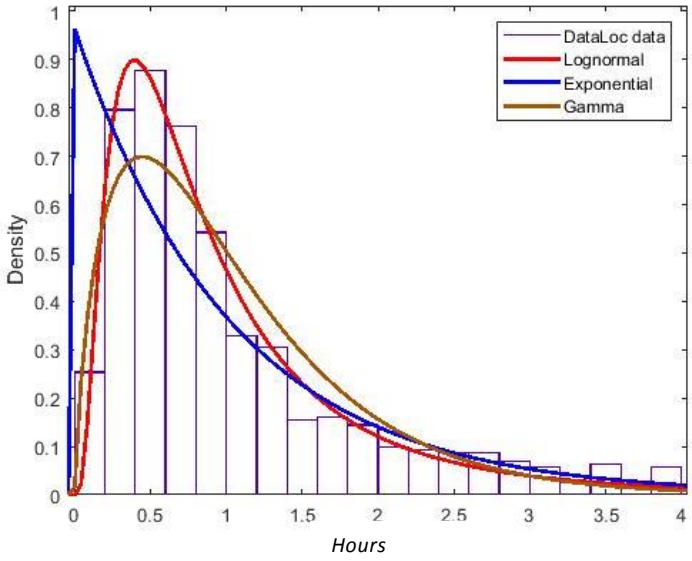


Figure 15 Histogram plotting local lead time data

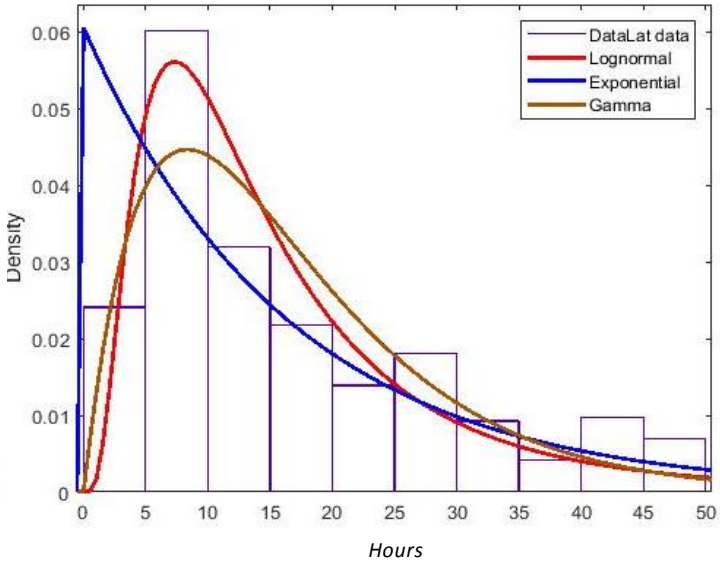


Figure 16 Histogram plotting lateral time data

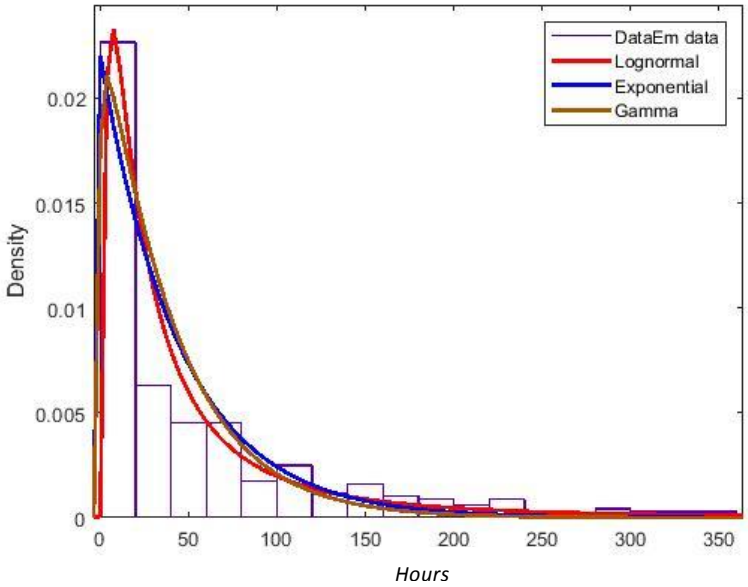


Figure 17 Histogram plotting emergency time data

Appendix E Warm-up times

We choose n = 5. This is the number of replications. Furthermore, we investigate when the SKU stock at the local warehouse in steady state. We took the time interval of 10 years and show the moving average over 5 weeks in Figure 18. As can be seen after 30 weeks the line is stabilized.

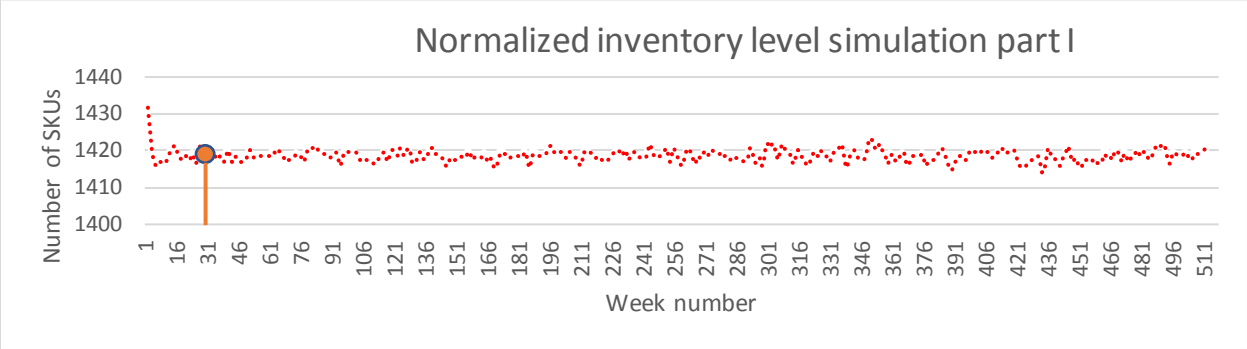
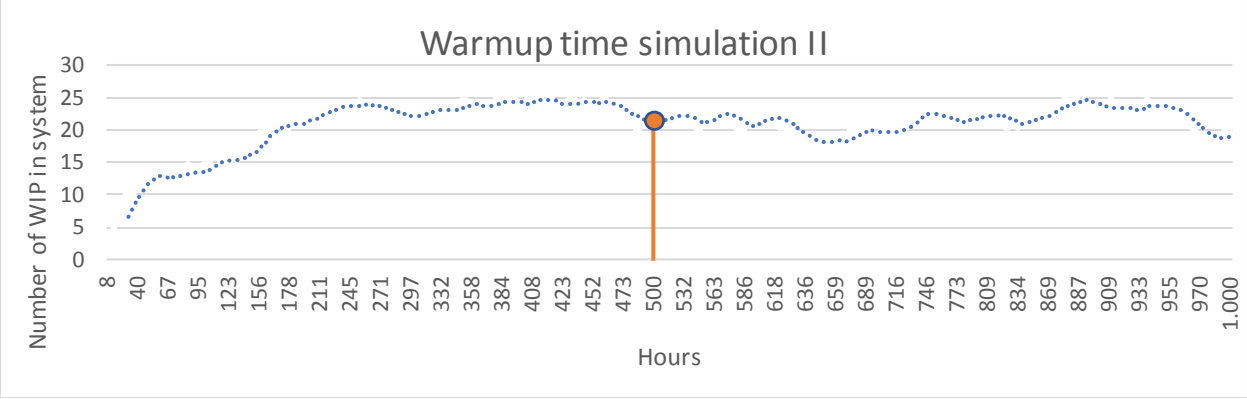


Figure 18 Warm up time simulation Part I

Figure 19 shows that around 250 hours the line is more or less stabilized. We take some margin and choose 500 hours to be the warm-up time. As parameters, we set the same as the simulation of Part I.

Figure 19 Warm up time simulation Part II



Appendix F Calculation of the number of replications

We calculate the number of replications or sub runs first for the simulation of Part I and then for the simulation of Part II.

Simulation I

By running the simulation model of part I, we obtain 40 sub runs so $q = 40$. Table 22 shows the results of these replications by using the formulas as provided in Section 4.3.2.

Table 22 Running simulation Part I

	Average (μ)	**Std. Dev (S)**	**CV**	**Minimal Q**
Fill rate	0.960	0.004	0.004	0.674
Mean waiting time for a spare part	0.010	0.001	0.095	86.853
Average Waiting time	3.1575	0.1613	0.0511	100.21

As can be seen is the waiting time value less accurate than the fill rate outcome. The minimal Q (replications) that is needed is rounded up, which results in 101 replications.

Simulation II

We run the simulation with 30 sub runs in order to find the minimal number of replications that are required. Since the CT and the WIP level are very dependent on the repair process, their variability is high. Therefore, the number of replications for a 1% deviation of the mean (w) very high. We choose to set this on 10% since the time to run the single scenario is 0.7 hour. For six scenario's this means that the total run time is 8 days.

Table 23 Running simulation Part II

	Average	**Variation**	**Std dev**	**CV**	**Q, w = .01**	**Q, w = .05**	**Q, w = .07**	**Q, w = .10**
Availability	0.804	0.000	0.009	0.011	5	1	1	1
CT	72.657	623.906	24.978	0.344	4541	196	100	46
TH	0.531	0.000	0.002	0.004	1	1	1	1
Utilization	0.779	0.000	0.008	0.011	5	1	1	1
WIP	38.534	171.500	13.096	0.340	4438	196	100	45

Appendix G Validation

We validate the simulation model based on the results of the tool SPartAn. Table 24 shows for every scenario the results and the deviation from the values by applying exponentially distributed lead times. Table 25 shows the results with distributions that we found in practice. We only consider the 'a' scenarios since the other scenarios cannot be compared with SpartAn on account of emergency times that are equal to the replenishment lead times.

Table 24 Validation simulation model exponential distributed lead times

Scenario	Serv. Meas.	Values spartan	Values simulation model	Difference
1a	Mean waiting time	1.00%	0.99%	1.00%
	Fill rate	96.02%	95.95%	0.07%
2a	Mean waiting time local	1.00%	1.00%	0.00%
	Fill rate local	93.13%	92.30%	0.89%
	Mean waiting time Main	1.00%	1.01%	1.00%
	Fill rate Main	96.04%	96.09%	0.05%
3a	Mean waiting time Loc	1.00%	1.00%	0.00%
	Fill rate Loc	93.68%	93.54%	0.15%
	Mean waiting time Main	1.07%	1.08%	0.93%
	Fill rate Main	95.52%	95.55%	0.03%

Table 25 Validation simulation model practical lead times

Scenario	Serv. Meas.	Values spartan	Values simulation model	Difference
1a	Mean waiting time	1.00%	1.00%	0.00%
	Fill rate	96.02%	96.80%	0.81%
2a	Mean waiting time local	1.00%	1.00%	0.00%
	Fill rate local	93.13%	92.63%	0.54%
	Mean waiting time Main	1.00%	0.99%	1.00%
	Fill rate Main	96.04%	96.11%	0.07%
3a	Mean waiting time Loc	1.00%	1.01%	1.00%
	Fill rate Loc	93.68%	93.96%	0.30%
	Mean waiting time Main	1.07%	1.06%	0.93%
	Fill rate Main	95.52%	95.61%	0.09%

Appendix H Input extreme values simulation models

Simulation Part I

In order to test the first simulation, we raised the base stock levels to 10 for every part. This is extremely high and must result in a low waiting time for a spare part and a fill rate of 1%. Table 26 shows the results of the simulation model. We conclude that the results are as expected and therefore the simulation model works as it should.

Table 26 Output simulation part I by extreme values

Fill rate	100%
Waiting time for a spare part	0.37%
Number of emergency transshipments	0
Number of lateral transshipments	0

Simulation Part II

In order to prove that by input extreme values the simulation still works as it should, we input the arrival of all wafer types at 10 batches in an hour. This is much more than the bottleneck process can produce in an hour. Furthermore, we use the values of scenario 1a that leads to the following results:

Table 27 Extreme values output

Description	**Value**
Average utilization	81%
Average WIP (units)	9.18e+04
Average throughput (units)	0.53
Average cycle time (hrs)	1.69e+05
Average availability	81%

Based on Table 27, we note that the utilization is almost equal to the availability. This is what we expected since the utilization is only determined of the time to repair of machines since the process is ove rloaded with wafers. Moreover, we note that the average WIP and cycle time are extremely high. Figure 20 shows that the WIP level in the system is increasing over time since in this example the process is unbalanced. This was what we expected, so we conclude that the simulation model works well.

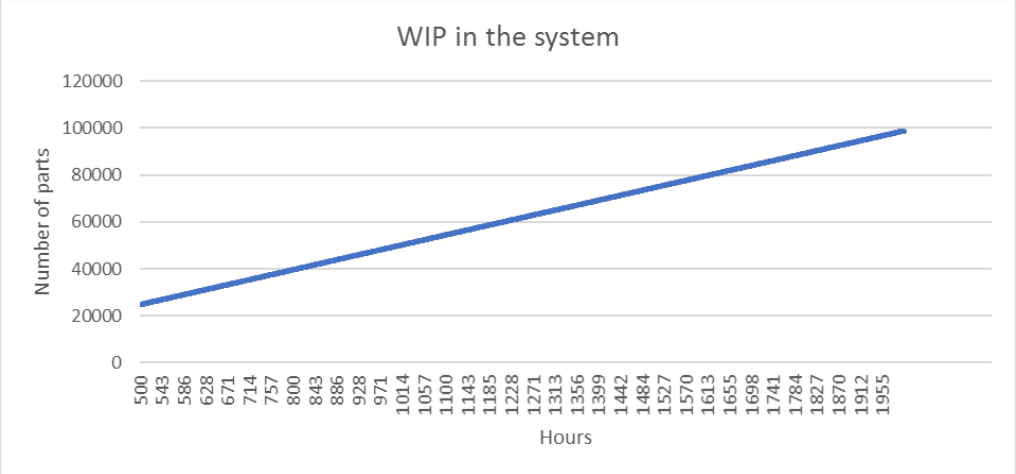


Figure 20 WIP level in the process

Appendix I Comparison of the number of machines at the customer

Figure 21 shows the fill rate level to obtain a mean waiting time for a spare part level of 1% at multiple machines (Z) at the customer.

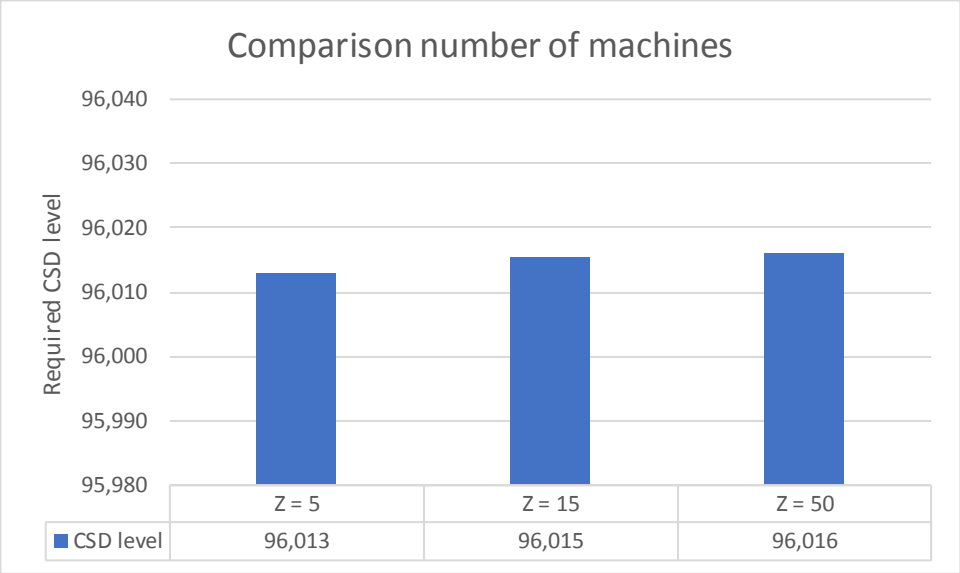


Figure 21 Comparison number of machines

Appendix J Waiting time probability distributions

All waiting time distributions that result from part I are given in the following figures:

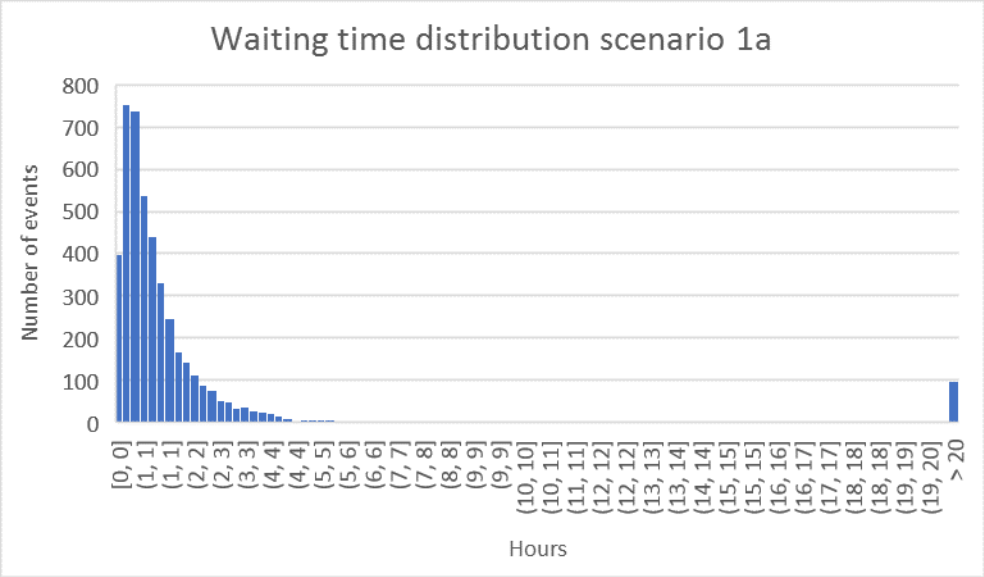


Figure 22 Waiting time distribution scenario 1a

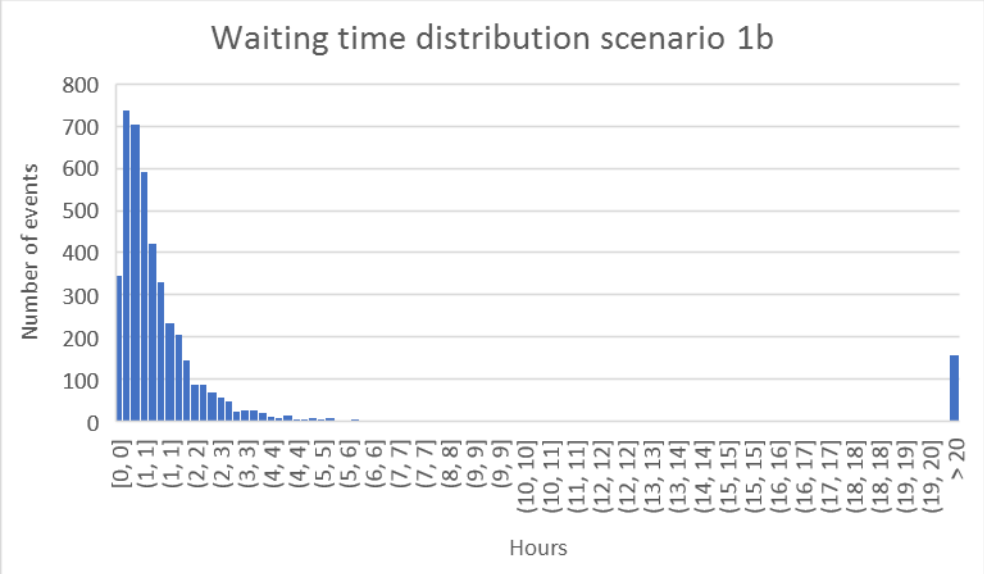


Figure 23 Waiting time distribution scenario 1b

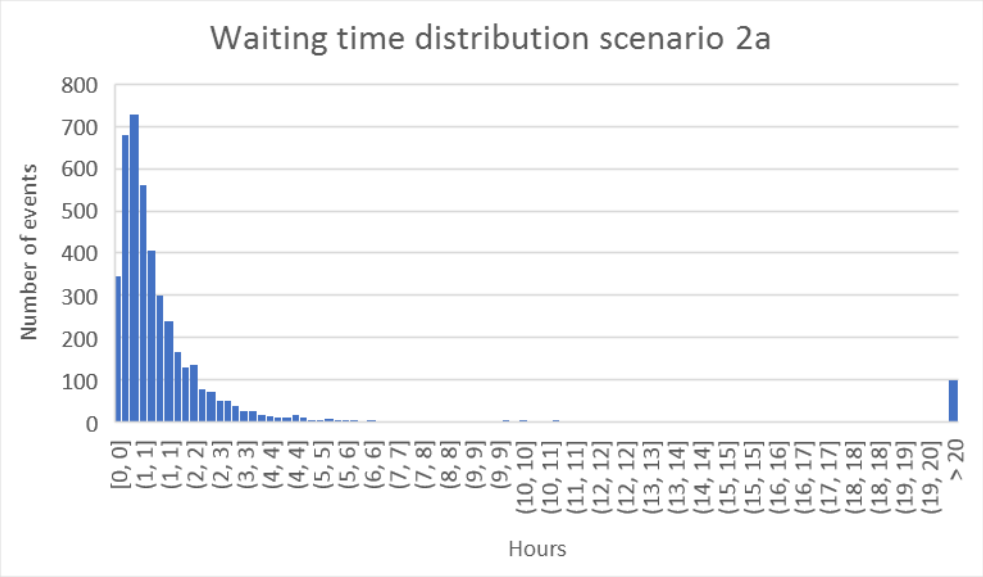


Figure 24 Waiting time distribution scenario 2a

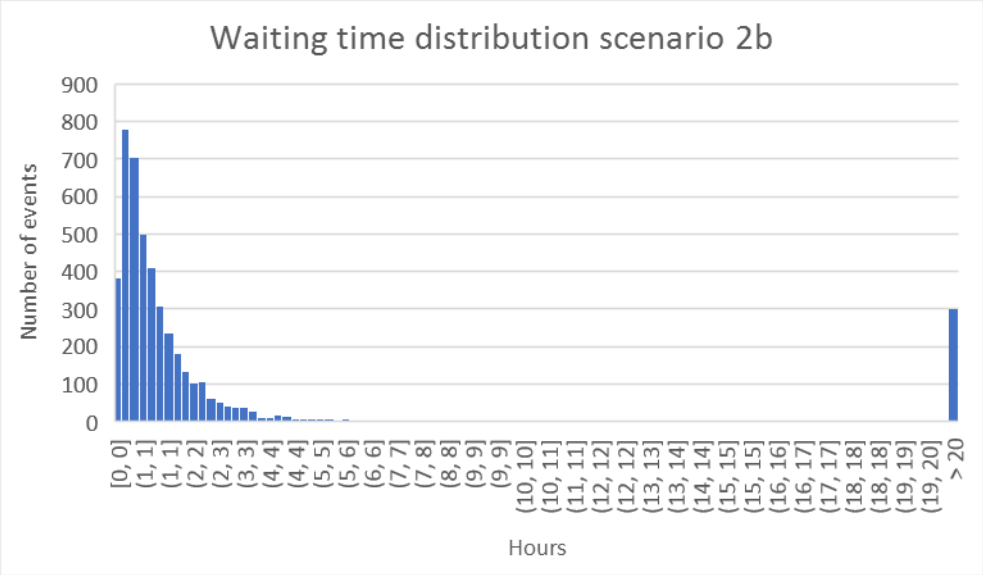


Figure 25 Waiting time distribution scenario 2b

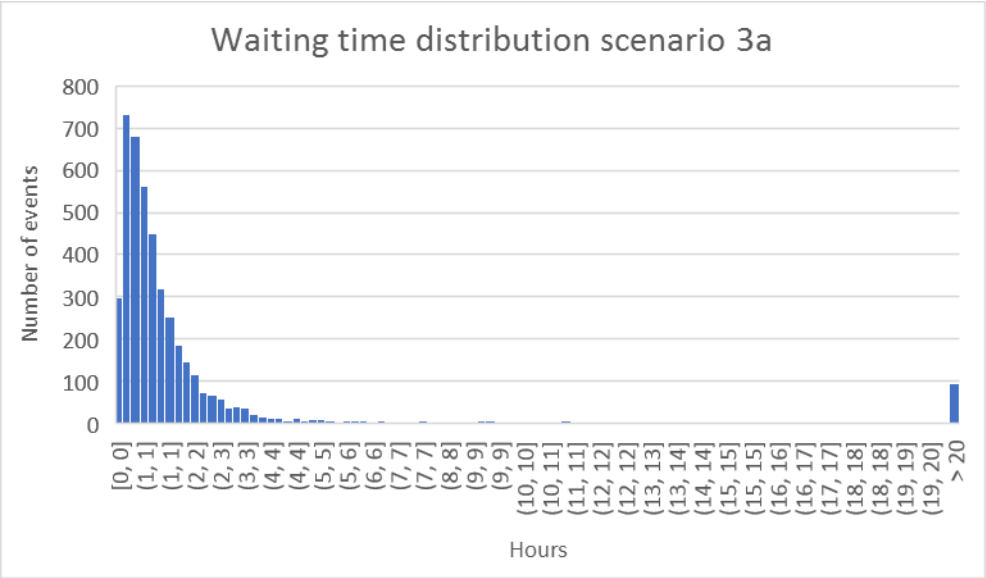


Figure 26 Waiting time distribution scenario 3a

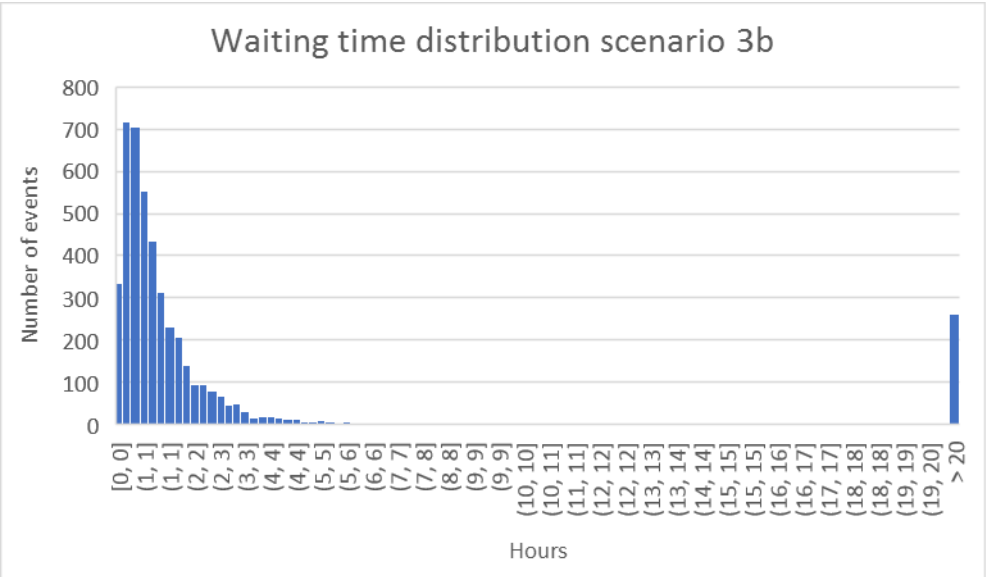


Figure 27 Waiting time distribution scenario 3b

Appendix K Transition matrix Repair process

The repair process can be given in a transition matrix as shown below:

$$Q = \begin{pmatrix} -15\lambda & 15\lambda & & & & & & \\ \mu & -(\mu+14\lambda) & 14\lambda & \cdots & & & & \\ & 2\mu & -(2\mu+13\lambda) & & & & & \\ & \vdots & & \ddots & & \vdots & & \\ & & & & -(4\mu+2\lambda) & 2\lambda & & \\ & & & \cdots & 4\mu & -(4\mu+2\lambda) & \lambda & \\ & & & & & 4\mu & -(4\mu+\lambda) & \end{pmatrix}$$

Q is a matrix with dimensions 15-by-15. The phases are noted as the number of broken machines. The rows denote the outflow and the columns the inflow in the particular phase.

Appendix L Probabilities to be in which phase of the repair process

Percentage of time, a certain phase occurs, is given in Table 28.

Table 28 Probability to be in a phase per scenario

SCENARIO	1A	1B	2A	2B	3A	3B
P0	5%	3%	5%	1%	5%	1%
P1	15%	11%	15%	4%	15%	6%
P2	24%	20%	24%	10%	24%	13%
P3	23%	22%	23%	15%	23%	18%
P4	15%	17%	15%	16%	15%	17%
P5	9%	12%	9%	15%	9%	15%
P6	5%	7%	5%	13%	5%	12%
P7	3%	4%	3%	11%	3%	8%
P8	1%	2%	1%	8%	1%	5%
P9	0%	1%	0%	5%	0%	3%
P10	0%	0%	0%	2%	0%	1%
P11	0%	0%	0%	1%	0%	1%
P12	0%	0%	0%	0%	0%	0%
P13	0%	0%	0%	0%	0%	0%
P14	0%	0%	0%	0%	0%	0%
P15	0%	0%	0%	0%	0%	0%

Appendix M Determine the input of the system

To determine the number of batches that arrive at the system, we use the solver of Microsoft Excel, as shown in Figure 28. As can be seen, the number of input has to be lower than the number of outputs. The solver gives the number of arrivals of a batch (red cells) as output.

Figure 28 Solvers excel

Scenarios 1a, 2a 3a				
INPUT				
	Number of arrivals	extra arrivals because of loop	Number of loops	Minimum number
A	0.095	2.00	21	0.01
B	0.116	2.20	19	0.01
C	0.073	1.46	20	0.01
D	0.116	2.20	19	0.01
E	0.132	2.77	21	0.01
OUTPUT				
Average processing time	1.1			
Average setups	0.15			
Lots per hour	0.90			
Machines	15			
Availability	0.8			
Lots per hour	11.20			
Total Input		Total output		
11.167	<	11.196		

Scenario 1b				
INPUT				
	Number of arrivals	extra arrivals because of loop	Number of loops	Minimum number
A	0.087	1.83	21	0.01
B	0.108	2.04	19	0.01
C	0.070	1.40	20	0.01
D	0.107	2.02	19	0.01
E	0.119	2.50	21	0.01
OUTPUT				
Average processing time	1.1			
Average setups	0.15			
Lots per hour	0.90			
Machines	15			
Availability	0.76			
Lots per hour	10.29			
Total Input		Total output		
10.28	<	10.29		

Scenario 2b				
INPUT				
	Number of arrivals	extra arrivals because of loop	Number of loops	Minimum number
A	0.079	1.66	21	0.01
B	0.094	1.79	19	0.01
C	0.064	1.28	20	0.01
D	0.095	1.81	19	0.01
E	0.101	2.12	21	0.01
OUTPUT				
Average processing time	1.1			
Average setups	0.15			
Lots per hour	0.9			
Machines	15			
Availability	0.67			
Lots per hour	9.09			
Total Input		Total output		
9.08	<	9.09		

Scenario 3b				
INPUT				
	Number of arrivals	extra arrivals because of loop	Number of loops	Minimum number
A	0.107	2.25	21	0.01
B	0.129	2.45	19	0.01
C	0.087	1.74	20	0.01
D	0.128	2.43	19	0.01
E	0.010	0.21	21	0.01
OUTPUT				
Average processing time	1.1			
Average setups	0.15			
Lots per hour	0.9			
Machines	15			
Availability	0.70			
Lots per hour	9.55			
Total Input		Total output		
9.54	<	9.55		