UNIVERSITY OF TWENTE.

MSc thesis

Improving the Service Control Tower with Proactive Decision Making

A Two-Staged Alert and Intervention Generating Model for Proactive Lateral Transshipments in an After-Sales Supply Chain

Ite Jan Muller

Public version

- version for publication -

October 25, 2018

Author:	I.J.T. Muller
Student number:	[redacted]
Contact:	i.j.t.muller@student.utwente.nl [redacted]

Company supervisor:	[redacted] Exception handling analyst
Contact:	[redacted]

ASML

ASML Netherlands BV
De Run 6501
5504 DR, Veldhoven

First supervisor:	Dr. M.C. Van der Heijden Associate Professor
Contact:	m.c.vanderheijden@utwente.nl
Second supervisor:	Dr. E. Topan Assistant Professor
Contact:	e.topan@utwente.nl

UNIVERSITY OF TWENTE.

University of Twente
Drienerlolaan 5
7522 NB, Enschede

Glossary

12NC Unique 12 digit material number for a Stock Keeping Unit. ii, v, viii, x, xiv, 9, 11–15, 23, 31–33, 35, 36, 38, 42, 43, 47–57, 60, 62, 68

Base Stock Level Order up to level of an item in the local warehouses and central warehouse. Also denoted as s in the $(s-1, s)$ policy notation, which is the replenishment and order policy at ASML. ii, vi, 4, 5, 14, 15, 25, 38, 68

Business Intelligence Technologies, applications and practices for the collection, integration, analysis, and presentation of business information. The purpose of Business Intelligence is to support better business decision making(OLAP Solutions, n.d.). ii, vi, 36

Confirmed usage Either usage from a sales order or usage of a part from a service event, where the part is actually used to replace a defect part in the machine. ii, 12, 17

Critical shortage Sum of the shortages (base stock level - regional inventory position) at the regions requiring support. ii, 33, 35, 37, 44, 45

Dead On Arrival A malfunctioning service part is shipped to the customer for a Service event. ii, vi

Deep Ultra Violet light Technology and business line of the previous generation Twinscan machines, capable of creating lines from 100 to 38 nanometer. ii, vi

Demand & Allocation Lead Function with the Field Material Availability department responsible for the demand outlook and scheduling parts at local warehouses. Each business line has one or more "DAL'ers"dedicated to their own region(s). ii, vi

Downtime Waiting for Parts Important performance indicator used as a target agreed with the customer in the Service Level Agreement (SLA). Defines how long the machine may be down waiting for a service part needed in the repair operation. ii, iv, vi, 5, 7, 29

Emergency Shipment type, fastest and most expensive shipment type. Duration no longer than 48 hours.. ii, 4, 29, 30, 56, 57, 62

Engineering Stock a.k.a. (Customer) Consignment Stock, is physically located at the customer site. A part is moved from the local warehouse to this location, before it is being transported to the machine in the cleanroom where it is needed.. ii, 4, 17

Extreme Ultra Violet light Technology and business line of the most recent machines, capable of creating lines smaller than 20 nanometer. ii, vi, 2

Field Service Defect Part that needs to be re-qualified, cleaned and/or repaired at the supplier. ii, vi, 4

Fill-rate Indicates the ratio of on-hand stock versus the base stock level. This is calculated for both the local warehouses as well as regions. ii, 8, 14, 16, 18, 31, 33, 68

Inventory position On-hand inventory minus backorders plus the pipeline inventory. ii, 23

Lateral transshipment “Lateral transshipments within an inventory system are stock movements between locations of the same echelon. ”(Paterson, Kiesmüller, Teunter, & Glazebrook, 2011). ii, 30, 57, 59, 60, 62

Lumpy demand Random demand with zero demand in many time periods but significant demands if other demand values can vary greatly (Regattieri, Gamberi, Gamberini, & Manzini, 2005). ii

Network Oriented Replenishment Application Automation tool built by ASML developers in the Excellence and Supply Chain Automation team within Service Management. Based on certain decision rules and the current inventory balance in the network, the tool creates purchase orders in SAP automatically. ii, vi, 5, 15, 23

Non-availability The part is not available to ship from the dedicated local warehouse to the customer upon request. Therefore an emergency shipment from another local warehouse or the central warehouse is needed. ii, vi, 3, 6, 30

Parts Request Form Process of requesting a part for a service order from the local warehouse. When the part is not available at the local warehouse, the part is sourced regionally or globally or eventually escalated to worldwide support in Veldhoven. ii, vi

Predecessor Older version of the service part. Sometime it is possible to use this part when the successor is requested. ii, v

Priority Shipment type, faster than routine with a duration no longer than 7 days. ii, 4, 29, 30, 56, 62

Re-inning The process of adjusting an existing purchase order or repair order with an earlier request date. ii, 15, 16, 68

Regional inventory position Inventory in the considered region (sum of inventories at local warehouses in the region + pipelines in between) minus backlog plus amount on order (in pipeline towards the region). ii, 36, 47, 60

Root Cause Analysis Systematic process for identifying root causes of problems or events and an approach for responding to them (Washington State Department Of Enterprise Services, n.d.). ii, vi, 15, 16

Routine Cheapest shipment type, with a duration no longer than 14 days. ii, 4, 29, 30, 62

Rush order Order from a local warehouse to the central warehouse to fulfill customer demand. The order must be linked to a service order. The fulfillment of the order from the central warehouse has priority over sales orders and normal replenishment. ii

Service event One of the following operations performed on a machine: Repair, maintenance, upgrade, installation of a new machine, or relocation of an existing machine. Repair and maintenance are part of the After Sales events. Upgrade, Installs and Re-locations are part of the UI& R event types. ii, iii, v

Service Level Agreement Contract between the customer and ASML which states the allowable Downtime Waiting for Parts (DTWP) when the machine is down. ii, iii, vii, 5, 29, 61

Single Point Of Contact Person who is responsible for all communication with a local. The responsibilities are divided per region. ii, vii

SPartAn Lost-sales based optimization algorithm to calculate order-up to levels of a part at the local warehouses. The network according to the algorithm consists of one central warehouse, regional and local warehouses. The process is executed periodically and reviews on part level are made in between. ii, 4, 57

Stock purge Event where some or all available stock of a part is being recalled due to a quality issue with the part. ii, 17, 19

Storage Location Administrative code for an item stored at a local warehouse. Indicates if the part can be used freely, or if it is restricted for usage, due to e.g. a reservation or quality issue. ii, vii

Successor New version of a service part replacing the Predecessor. This part has a different 12NC code, often a one number increment of the previous 12NC. ii, 17

Supply Chain Visibility "The ability to be alerted to exceptions in supply chain execution (sense), and enable action based on this information (respond). In essence, visibility is a sense and respond system for the supply chain based on what is important in the business" (McCrea, 2005). ii, vii, 5, 10, 11, 18, 30

Upgrades, Installs and Re-locations Department within Supply Chain Management and Service event type. ii, vii

Usage for analysis Using a part during a repair operation to see if the part solves an issue with the machine. When the part is not needed after the analysis, it has to be re-qualified and repacked, before it can be used again. Generally, this needs to be done at the supplier of the part. ii

Acronyms

ADI Average Demand Interval. ii

BI Business Intelligence. ii, vi, 11, 32, 36, *Glossary:* Business Intelligence

BSL Base Stock Level. ii, vi, 4, 5, 14–16, 18, 25, 30, 38, 47, 53, 54, 56–58, 68, *Glossary:* Base Stock Level

CBM Condition Based Maintenance. ii, 21, 22

CPA Continental Parts Availability. ii

CSCM Customer Supply Chain Management. ii, 2, 5, 11, 14, 23, 25, 33, 37, 42, 58

CV Coefficient of Variation. ii

DAL Demand & Allocation Lead. ii, vi, 58, 60, *Glossary:* Demand & Allocation Lead

DOA Dead On Arrival. ii, vi, 17, 19, *Glossary:* Dead On Arrival

DTWP Downtime Waiting for Parts. ii, iv, vi, 5, 7, 8, 29, *Glossary:* Downtime Waiting for Parts

DUV Deep Ultra Violet light. ii, vi, 54, *Glossary:* Deep Ultra Violet light

EOL End of life. ii

ETA Estimated Time of Arrival. ii, 14, 19, 26

EUV Extreme Ultra Violet light. ii, vi, 2, 54, 58, *Glossary:* Extreme Ultra Violet light

FMA Field Material Availability. ii, 42, 53, 58

FSD Field Service Defect. ii, vi, 4, 5, *Glossary:* Field Service Defect

GPA Global Parts Availability. ii

LPA Local Parts Availability. ii, 7, 54

NAV Non-availability. ii, vi, xiv, 3, 6–8, 30–33, 52, 54, 56, 57, *Glossary:* Non-availability

NORA Network Oriented Replenishment Application. ii, vi, 5, 11, 15, 23, 26, 28, 32, 35, 36, 38, 53, 62, 63, 69, *Glossary:* Network Oriented Replenishment Application

PRF Parts Request Form. ii, vi, *Glossary:* Parts Request Form

RCA Root Cause Analysis. ii, vi, 15, 16, *Glossary:* Root Cause Analysis

ROP Reorder Point. ii, 4

SCV Supply Chain Visibility. ii, vii, 5, 10, 11, 18, 20, 21, 23, 30, 59, 61, 62, *Glossary:* Supply Chain Visibility

SLA Service Level Agreement. ii, iii, vii, 5, 25, 29, 61, 62, *Glossary:* Service Level Agreement

SLOC Storage Location. ii, vii, 19, *Glossary:* Storage Location

SPOC Single Point Of Contact. ii, vii, *Glossary:* Single Point Of Contact

SSL Safety Stock Level. ii, 4

UI& R Upgrades, Installs and Re-locations. ii, vii, *Glossary:* Upgrades, Installs and Re-locations

Symbols

Variables

AGR Actual goods receipt quantity

BSL Base stock level

$\mathbb{E}\{NAV\}$ Expected unplanned non-availabilities

L^{NB} New Buy Lead Time

L^{R} Replenishment Lead Time between central warehouse and local warehouse

λ Forecast

OH On-hand inventory level

IP Inventory Position

SR Scheduled receipt quantity

U Confirmed usage

SN Support Needed ($\in \mathbb{B}$))

SP Support Possible ($\in \mathbb{B}$))

$x_{a,r,c}$ Decision variable in the intervention generating model. Determines the optimal Inventory Position change as a result of the intervention

Sets

I All 12NC stock keeping units

A Set of 12NCs for which an alert is generated

J All warehouses, where j=0 denotes the Central Warehouse

R All regions

Executive Summary

The following graduation thesis *Improving the Service Control Tower with Proactive Decision Making*, is about introducing proactive lateral transshipments to the After-Sales Service Supply Chain of ASML. ASML is one of the world's leading manufacturers of chip-making equipment. ASML develops and manufactures high tech lithography machines used to produce semi-conductors. As well as other capital goods manufacturers, ASML offers after-sales service to facilitate the customers demand for high availability of the machines. ASML uses a global network of warehouses stocking spare parts, to guarantee the promised response time on machine failure.

ASML refers to the Control Tower process as a process that recognizes exceptions in the Supply Chain network and initiates interventions to avoid a non-availability. A non-availability occurs when a part is needed however, not available at the right warehouse. The current Control Tower dashboard contains data on inventory, supply, usage and forecast and is capable of generating alerts for situations marked as an exception.

The goal of this thesis is to enable proactive decision making in the Control Tower process, so that more non-availabilities can be avoided. The research question is therefore formulated as: How can prospective unplanned non-availability issues of after-sales service parts be recognized, analyzed and pro-actively acted upon, to increase the spare part availability of these parts on an operational level?

At the beginning of the project, I have analyzed the current decision rules in the Control Tower and how the alerts are processed. Improvement potential for the existing Control Tower are 1. re-examining the parameters that trigger an alert, 2. more proactive intervention suggestions, and 3. using information at a lower aggregation level to identify prospective non-availabilities.

The proactive operational interventions that I have found in literature are: 1. proactive stock allocation, 2. proactively skipping replenishments, 3. dynamic stock rationing, 4. proactive emergency shipments from an upstream warehouse, and 5. proactive lateral transshipments. Of these interventions only the last two are actually suitable for the ASML supply chain. I chose to focus the remainder of the research on implementing Proactive Lateral Transshipments to be executed when the Central Warehouse is unable to replenish the shortages at the local warehouses. The options for a proactive intervention on supply side are limited. In these situations, proactive lateral transshipments can minimize the risk on a non-availability with the available resources by balancing the risks in the network.

I chose to focus on transshipping parts between regions instead of between all local warehouses so that a reduced optimization problem remains (■ regions instead of approximately ■ local warehouses). When a lateral transshipment between two different regions is proposed, a planner should determine from and to which local warehouse the part is sent. Another reason is that the calculation for expected unplanned non-availabilities in the next month for a local warehouse only includes standard replenishment lead time and does not include the reactive transshipment option that is used within ASML. When a local warehouse is out-of-stock, another local warehouse in the same region can support quickly. This risk pooling effect is included in the calculation of expected unplanned non-availabilities in the next month in a region. Therefore the reduction of non-availability risk by using proactive lateral transshipments is much higher for lateral transshipments between regions than lateral transshipments between local warehouses.

To propose the proactive lateral transshipment intervention in a certain situation, I have developed a two-staged model. The first stage of the model recognizes the 12NCs with an unbalanced situation and generates an alert. The second stage of the model generates the corresponding intervention to balance the risk in the regions. I chose this two stage approach since optimizing the network for the tens of thousands unique stock keeping units can take up to a day of running time. Therefore the alert generating stage makes a pre-selection of reasonable alerts.

The risk on a non-availability is used as the measure that needs to be balanced between regions. ASML expresses this risk as the expected unplanned non-availabilities in the next month.

Alert generating model I have developed decision rules that together generate an intervention when: 1. The Central Warehouse has zero on-hand inventory; 2. The Scheduled Receipts quantity is less than ■ % of the shortages in the regions requiring support; 3. There is at least one region that requires support (with an unplanned non-availability risk higher than ■ expected unplanned non-availabilities in the next month); 4. There is at least one region that is able to support (with an unplanned non-availability risk lower than ■ expected unplanned non-availabilities in the next month) and an inventory position of ■ parts or more[1].

Intervention generating model Next, I developed a decision support model that generates intervention proposals to optimally redistribute stock between regions. This intervention generating model proposes an intervention for individual parts with an alert by solving a linear programming model. The objective function is to minimize the maximum regional expected unplanned non-availabilities, such that the risk on a non-availability is balanced between the regions. Constraints to the model are: 1. The Regional Inventory Position after the intervention is the initial Inventory Position plus or minus the proposed change; 2. The quantity of the parts sent must be equal to the quantity of parts received; 3. A decision on the Inventory Position of a region after the interventions must be made; 4. The Inventory Position can only decrease in a region when the region is eligible to *give* support; 5. The Inventory Position can only increase in a region when it is eligible to *receive* support; 6. At least one part remains in the supporting region.

Numerical results The number of parts for which an intervention is justified is around■ parts per week. With the proactive lateral transshipments, the number of new alerts per week will decrease with on average 53%, assuming that an intervention is sufficient to resolve the imbalanced network such that the alert is removed the next week. The impact of the proactive transshipment on the imbalanced network is a reduction of the maximum expected unplanned non-availabilities in the next month of 84% on average.

On a yearly basis, approximately ■ unplanned non-availabilities will be avoided by proactive lateral transshipments. Approximately ■ inter-regional emergency shipments will be avoided by the proactive lateral transshipments. These emergencies can be requested by a region for various reasons. The costs of the transshipments outweigh resolving downtime caused by a non-availability. For one hour of downtime saved, around ■ priority or ■ routine proactive lateral transshipments can be used [2].

Recommendations As a suggestion for further research, I recommend a simulation study which can further support decision making and improvement of the decision rules. With respect to the implementation of this thesis at ASML, I recommend to use the accompanying tool once a week and execute the proposed interventions. It is recommended to structurally measure the impact after the intervention is executed that enables also Machine Learning in the future. Furthermore, it is recommended to share historical data on shipments with the whole organization such that business decisions can be made with reliable data on the costs of the decision.

[1] The parameters are removed in this version for publication.

[2] The exact numbers are removed in the version for publication. It can be said that the costs of downtime are dis-proportionally higher than the costs of a transshipment

Preface

Dear reader,

Here it is! My final thesis, describing the research I did on the subject of Service Control Towers. Before I started, I was looking for a graduation project involving Supply Chain Management at a large company. Thanks to a message from my supervisor, Matthieu van der Heijden, this search ended successfully with an assignment at ASML. I have really enjoyed my time at ASML and especially in the Service Management team. I believe I have now gained more knowledge in the subject of After-Sales Service Supply Chains and Business Intelligence. Furthermore, I have now experience working at an international high-tech firm, which has its own perks and drawbacks.

I would like to thank my supervisors Matthieu van der Heijden and Engin Topan at the University of Twente for their support and critical feedback. More than once you have focussed my activities and narrowed down the scope. Your extreme eye for detail and challenging questions have helped me to improve my reporting skills and focus on what is really important.

Furthermore, I'd like to thank Jacky van de Griendt and Ruud van Sommeren for giving me the opportunity and supporting my work. I feel proud having contributed to the Service Management team and having the tool I built really implemented. This would not have been possible without your trust and support. Furthermore, it has been really helpful that I could always schedule a meeting with Jacky when I needed to get my thoughts straight. Thanks to our sparring sessions, I was able to proceed with new ideas whenever I got stuck.

A special thanks goes out to all the people in the Service Management team. I have felt very welcome and really a part of the team. Also thanks to all your individual expertise I was able to quickly grasps the complex structure of ASML's Supply Chain. But above all, it's been fun!

Lastly, I would like to thank my dear family and girlfriend without whose motivation and support, I would never have been able to finish this assignment. Quite literally since my grandmother could arrange a room for me to stay during the last phase of the research, in which I had to leave my apartment. A last word of thanks goes out to my car, that so far has brought me back and forth between Veldhoven and Stadskanaal every weekend.

For now, I can only hope that in this thesis, you can recognize the effort, joy and personal development I have experienced during the execution of this project.

Ite Jan Muller
Stadskanaal, October 25, 2018

Contents

List of Figures

List of Tables

List of Decision rules

ASML

It's not what you look at that matters, it's what you see.

Henry David Thoreau

UNIVERSITY OF TWENTE.

Chapter 1

Introduction

This thesis in front of you is the result of the master research to finish my Master Program Production and Logistics Management at the University of Twente. The research project is carried out at the Customer Supply Chain Management department of ASML in Veldhoven. At ASML, a Supply Chain Control Tower, using recent data about the status of the whole network, is used to track down exceptions or threatening availability issues. After analyzing, corrective or proactive interventions are proposed. This project focuses on improving the current Control Tower with proactive decision making and interventions to improve the availability of spare parts.

This chapter explains the background of this assignment by describing the company, the department and the control tower process.

1.1 Background

ASML is one of the worlds leading manufacturers of chip-making equipment. The headquarter is located in Veldhoven (the Netherlands), and the company employs over 19.000 people in 16 countries. The vision of ASML is "a world in which semiconductor technology is everywhere and helps to tackle societys toughest challenges. (...) We contribute to this goal by creating products and services that let our customers define the patterns that integrated circuits are made of, and we continuously raise the capabilities of our products, enabling our customers to increase the value and reduce the cost of chips."
"By helping to make chips cheaper and more powerful, we help to make semiconductor technology more attractive for a larger range of products and services, which in turn enables progress in fields such as healthcare, energy, mobility and entertainment"(ASML, 2018b).

ASML develops and manufactures high tech lithography machines used to produce semiconductors. Semiconductor chips are made on a silicon disk called a wafer. The lithography machine of ASML projects a pattern of small lines on a light sensitive layer that is applied to the wafer. After exposure to UV light, the wafer is developed and either the exposed or the unexposed part of the light sensitive layer is then removed from the wafer. Then the unprotected part is etched out of the wafer or a conducting material is applied to the wafer. These steps are repeated layer after layer and the wafer is sawn into small chips. The treatment of the wafer is shown in figure 1.1.

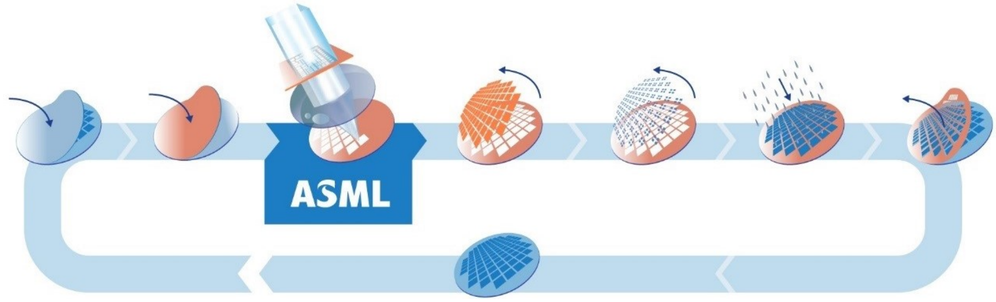

Figure 1.1: Wafer treatment process (ASML, 2017)

The products of ASML are divided in three systems: The mature PAS system, which is not produced anymore, but still being maintained and refurbished, and the TWINSCAN XT/NXT and TWINSCAN NXE systems. The XT/NXT machines use a light source with a higher wavelength. The newest NXE system uses Extreme Ultra Violet light (EUV), enabling the exposure of very small lines (between 30 nm and < 20 nm) on the wafer. Furthermore, ASML produces the YieldStar metrology systems to enhance the efficiency of the chipmaking process.

1.2 CSCM Service Management

As stated in the introduction of this section, the goal of this research is to improve ASML's Service Control Tower. The Service Management department owns the responsibilities for the execution of the Control Tower process. Service management is positioned in the organization as a support department to the Customer Supply Chain Management (CSCM) departments (Figure 1.2). The responsibilities regarding the Control Tower are located in one of the Service Management teams: The Report Industrialization and Control Tower Analytics team. This team is furthermore responsible for report creation and maintenance and tactical control tower analytics including trend analysis. The team is doing projects to structurally improve the customer supply chain and analyze exceptions based on the alerts that the control tower dashboard generates.

One of the other teams within Service Management is Service planning. Service Planning is responsible for setting the targets of service parts and tools in the central and local warehouses (tactical planning) and making the forecasts for service parts and tools. The Excellence & Supply Chain Innovations team focusses on automatization of the processes. BPM & FMO focus on the future mode of operations and are planning the transition of the whole organization to only use standard SAP transactions. The deal structuring and budget control team has just been established and is responsible for determining the supply chain related annual budgets.

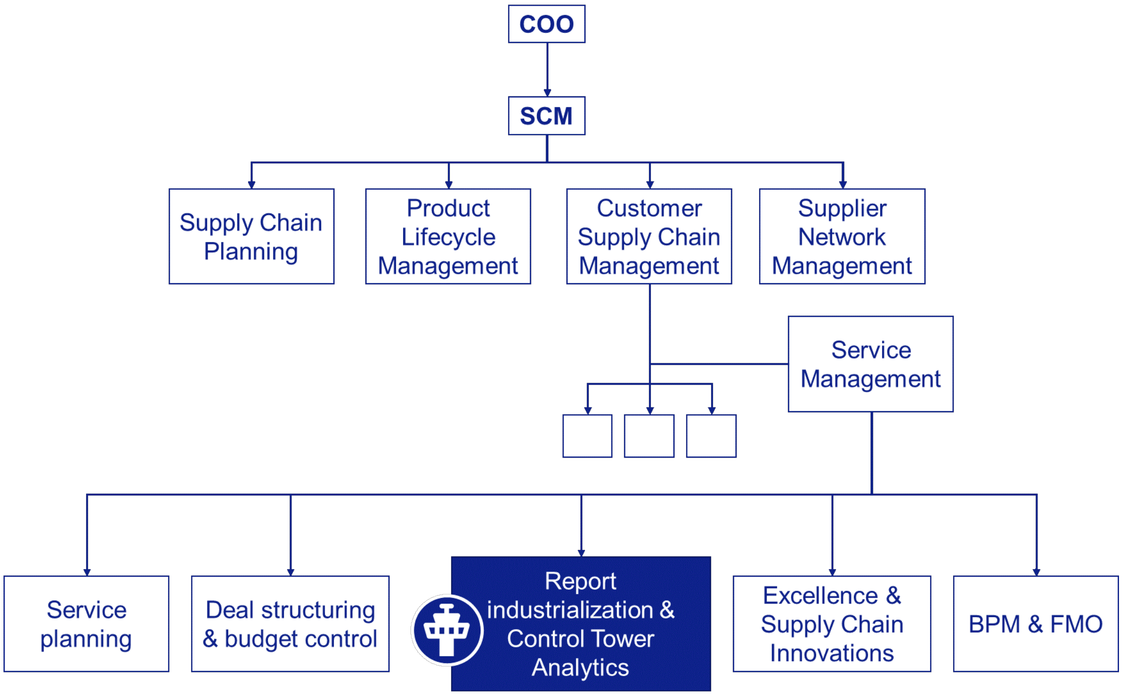

Figure 1.2: Organogram Supply Chain Management at ASML

1.3 ProSelo Next

This MSc. assignment is initiated by ASML because of their contribution to the ProSeLo Next consortium. ProSeLo Next stands for Proactive Service Logistics for Advanced Capital Goods the Next Steps. This consortium is a research agreement between four universities and nine companies: University of Twente, University of Tilburg, University of Technology Eindhoven, Erasmus University, ASML, Fokker Services, IBM, Marel Stork, Océ, Thales Nederland, Vanderlande, Gordian Logistic Experts, and the Stichting Service Logistics Forum. The project is split into three work packages: 1) Predictive maintenance and service logistics, 2) Service business models and 3) Service Control Towers(Basten, 2016). This project is part of the third work package.

1.4 Customers and Supply Chain

The customers of ASML are the worlds most leading chip manufacturers (e.g. Intel, Samsung and TSCM), who use the machines of ASML to make a wide range of semiconductor chips (ASML, 2018a).

The installed base of ASML machines is spread throughout the world, in the continents Asia, United States and Europe. ASML offers after-sales service to enable maximal utilization of the machines for the customers of ASML. Numerous (~ 50) local warehouses are located 'in the field' near the customers, to be able to quickly deliver service parts and tools when needed. A central warehouse replenishes these local warehouses. Another supply hub acts as a central warehouse for distinct service parts and acts as an emergency hub in case a local warehouse faces a Non-availability (NAV). A graphical representation can be found in figure 1.3.

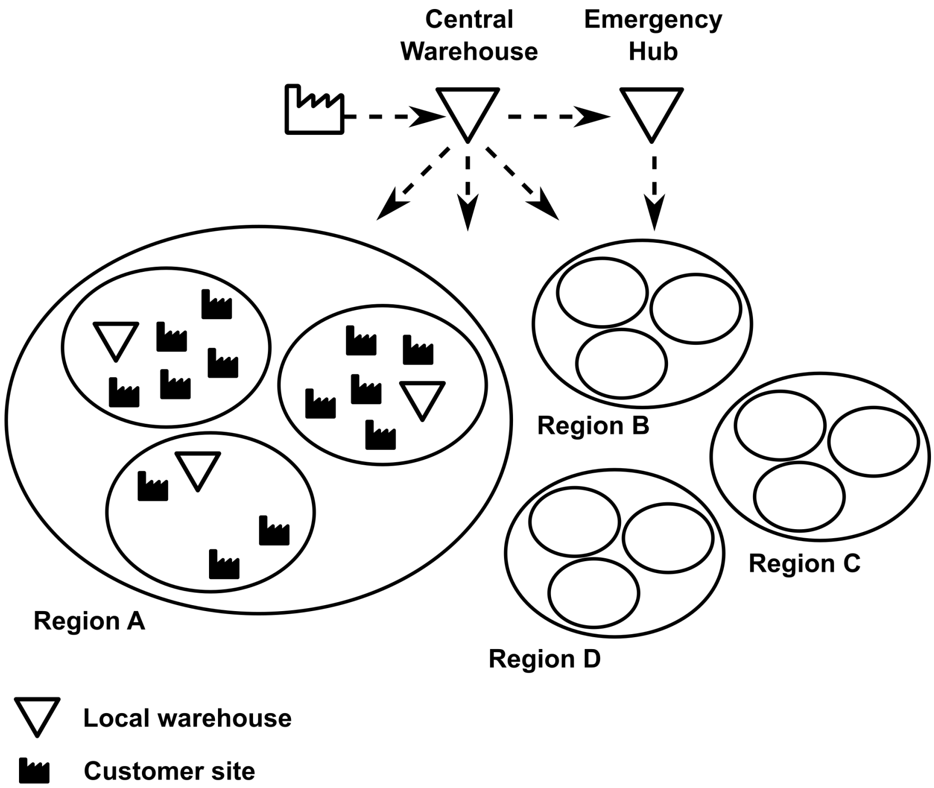

Figure 1.3: ASML supply chain network

In the event of a service, either by failure or maintenance, a part is removed from the machine and returned to the supplier for repair, cleaning and/or requalification. This is called the Field Service Defect (FSD) return flow. This flow is graphically represented in figure 1.4. The target setting of the Reorder Point (ROP) in the Central Warehouse includes the repair success rate of these parts and are the responsibility of the Service Management team.

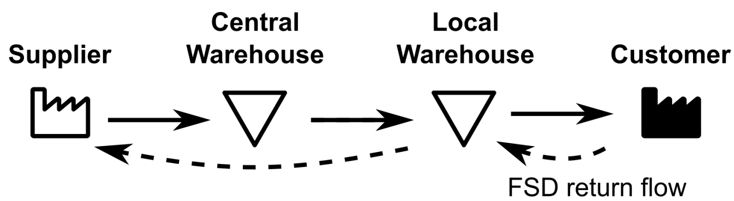

Figure 1.4: Repair flow in the supply chain of ASML

At ASML, an $(s-1, s)$ inventory policy is used for replenishing the Central and Local Warehouses. This means when a part is consumed from a local warehouse, a replenishment of the Local Warehouse is triggered. This will trigger replenishment from the Central Warehouse as well. At the Local Warehouse, the Base Stock Level (BSL) s is called a Safety Stock Level (SSL) in ASML (contrary to the definition in literature) which is determined by the SPartAn algorithm. At the Central Warehouse, the order-up-to level s is called a Reorder Point (ROP) and roughly equals the demand during lead time + safety stock level for demand during lead time. When a part is brought to the machine to be used for a service operation, it is moved administratively to so called Engineering Stock. If the part is not used during the operation or only used for the analysis of the problem, it is send back to the local warehouse. The part needs cleaning and/or re-qualification first before it can be used again.

ASML applies three shipment types for movements between the Central Warehouse and Local Warehouses: Routine, Priority and Emergency. A lateral transshipment is now only applied in case of emergency, in these case a Priority or Emergency shipment is used. Table 1.1 explains the differences. Please note that the costs are averages including high volume parts in the lowest weight category and emergency shipments within the regions. Therefore not accurate, but the best estimate possible. Furthermore the costs are normalized for confidentiality purposes.

Table 1.1: Shipment types

Shipment type	**Duration**	**Normalized costs**
Routine	< 14 days	1
Priority	< 7 days	1.5
Emergency	< 48 hours	4

The Excellence & Supply Chain Innovations team developed the Network Oriented Replenishment Application (NORA) tool, that analyzes the Supply Chain of ASML every morning. When the inventory position of a local warehouse is lower than the BSL, a replenishment shipment is scheduled automatically. When multiple warehouses have a shortage (if the BSL is higher than the inventory position) and need replenishment, a prioritization rule is used based on the risk on a non-availaiblity in that local warehouse. The priority is determined so that the first part is shipped to a local warehouse in the continent and region with the highest risk. The tool also triggers new-buy orders and the repair of FSDs

1.5 Service level agreements

The customers of ASML demand high availability of the lithography machines since the machine is the bottleneck in the chip making process. To ensure the availability level of the machine, ASML offers the customers an Service Level Agreement (SLA) in which a certain maximum Downtime Waiting for Parts (DTWP) in 13 week periods is promised. When the agreed availability target is not met, ASML faces dissatisfaction. Therefore, maximizing the availability of spare parts in the supply chain network is the most important target of the CSCM department. The target of the Control Tower is giving Supply Chain Visibility (SCV) to detect those situations where the availability of spare parts is at risk and to suggest an appropriate response (See section 2.3 for more information on SCV).

There are two levels of Service Level Agreements: Local and Non-local. The Local contracts promise the lowest DTWP per 13 weeks, meaning spare parts need to be located close to the customer site. The non-local contracts do not have this requirements, therefore when needed, a part can be shipped from another local warehouse in the continent or from the central warehouse. These service commitments are important input when determining the Base Stock Level (BSL) at the local warehouses.

1.6 Service Control Tower

The most referenced definition of a Supply Chain Control Tower uses five layers: "Supply Chain Business layer, Information Perception layer, Information Operational Control layer, Information Service Platform layer and the Information Manpower layer" (Shou-Wen, Ying, & Yang-Hua, 2013). In a master thesis by Schwagmeier (Schwagmeier, 2016), the theory of Supply Chain Control Towers is applied to the after-sales supply chain. Figure 1.5, adapted from this thesis, shows the Control Tower layers.

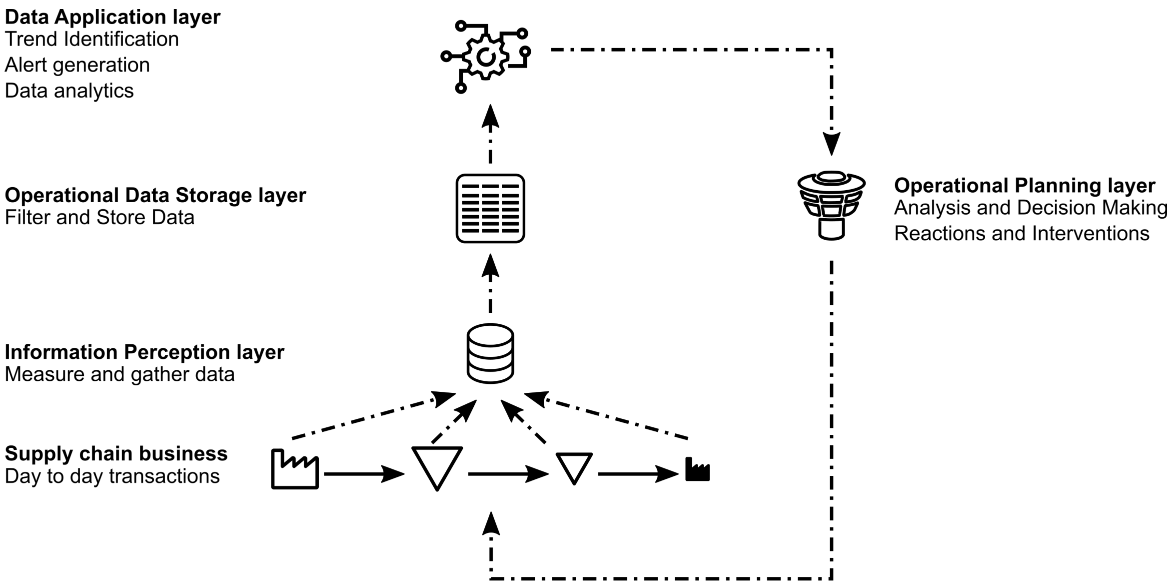

Figure 1.5: Five layers of the Control Tower (adapted from: (Schwagmeier, 2016))

The bottom-up structure converts real-time movements and events on the bottom layer to manageable information in the top layer. The information is presented using a dashboard to planning operators monitoring the Supply Chain. “This allows people trained to use these visibility capabilities to detect and act on risks or opportunities more quickly”(Shou-Wen et al., 2013).

The Control Tower at ASML is designed and implemented in ■ . A common used description for the Control Tower is: “a centralized hub that uses real-time data from a companys existing, integrated data management and transactional systems to integrate processes and tools across the end-to-end supply chain and drive business outcomes ”(Bleda, Martin, Narsana, & Jones, 2014).

A Service Control Tower can be defined as a Supply Chain Control Tower covering the after-sales service Supply Chain. This is the network enabling the delivery of spare parts to the customer.

1.7 Non-availabilities

A Non-availability (NAV) occurs when a part is not available at the local warehouse, when it is requested for an after-sales service order. The service management team distinguishes two different kinds of NAVs: planned and unplanned, that will be explained in this section.

1.7.1 Planned NAVs

As described earlier in section 1.5 , the base stock levels are determined by an optimization algorithm. The algorithm tries to find the best inventory policy such that the parts availability is as high as reasonable. Therefore, not every location will have a base stock level for a certain part. When there is a request for that part from that local warehouse, a non-availability occurs. These are so called planned NAVs.

1.7.2 Unplanned NAVs

An unplanned NAV occurs when there is a positive base stock level, but there was no part available from stock at the requested time. The goal of the Control Tower is to avoid unplanned NAVs as much as possible. The official KPI definition of an unplanned NAV is as follows:

This paragraph is removed in the version for publication.

1.7.3 Emergency handling

In case of a local non-availability on an LPA contracted service part, an emergency shipment of that part is requested. The first step is to look for an available part elsewhere in the region, such that the part can be shipped quickly to the customer. If that is not possible, sourcing is attempted sequentially from continental stock or worldwide stock, this might be from another local warehouse, the central warehouse, or the emergency hub.

1.8 Problem cluster

A non-availability impacts the customer and ASML since the customer loses valuable production output when a machine is Downtime Waiting for Parts (DTWP). To avoid DTWP, the Control Tower function tries to maximize material availability on an operational level by pro-active decision making and challenging the target settings and forecasts from the tactical level without overruling tactical levels/plans (too much).

When analyzing the current situation, I have found that the manual analysis of all the alerts that are generated by the Control Tower tool is an inefficient process that requires a lot of resources. Most of the generated alerts are classified as a *false alert*, by filtering the parts with an alert on some criteria. However, these filters let through these *false alerts* making the process inefficient.

This paragraph is removed in the version for publication.

When too few pro-active measures to avoid non-availabilities are carried out, eventually a non-availability will happen. This will result in risk of higher Downtime Waiting for Parts for the customer and higher emergency costs for ASML. The breakdown in table 1.1 shows that emergency shipments are approximately 4 times as expensive as normal shipments. More importantly, downtime leads to dissatisfaction from the customer. This problem is summarized in the problem cluster in figure 1.6.

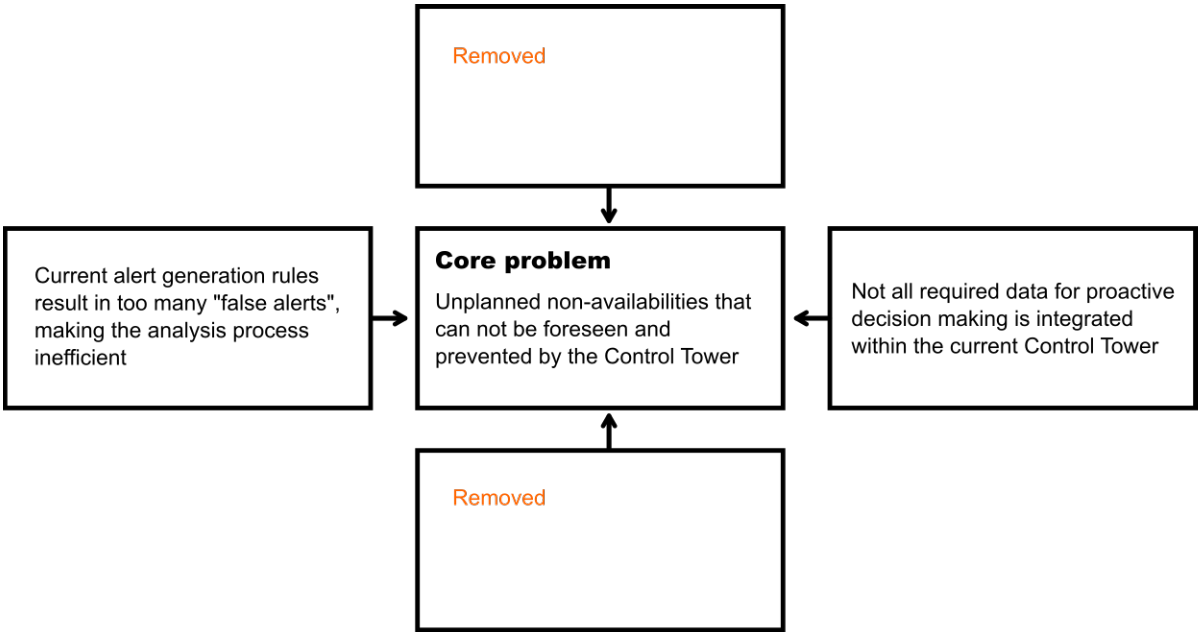

Figure 1.6: Problem cluster

1.9 Research goal

The goal of this research is to solve the action problem of ASML. The action problem is derived from the core problem (problem cluster) from the previous section and will be described in the following section. The research goal is described in the section thereafter.

1.9.1 Action problem

Unplanned NAVs impact the customer by increasing the expected DTWP. The priority of ASML and the Control tower is to maximize the availability of service parts for the customer, but NAVs will eventually lead to increased financial risk for ASML as well.

The Control Tower cannot prevent most of the unplanned non-availabilities because:

1. the non-availabilities cannot be foreseen with the current Control Tower alert triggers or
2. generated alerts are not analyzed and acted upon due to capacity issues within the Service Management team.

The variable for this problem is material availability. The indicators that are used or can be used in the future are the number of unplanned NAVs, emergency orders, Fill-rate, downtime risk and the number of alerts in the control tower. Part of the knowledge problem is to find out what figures indicate an upcoming issue in the supply chain. Therefore, this list of indicators is incomplete and will be extended during the project.

The problem owner is the student performing this research. Stakeholders are the Control Tower team within the Service Management team and eventually the customers of ASML when material availability is improved. From this action problem, the research goal can be defined. The research goal is described as the main research question and the solution to this question should solve the action problem stated earlier.

1.9.2 Research goal

How can prospective unplanned non-availability issues of after-sales service parts be recognized, analyzed and pro-actively acted upon, to increase the spare part availability of these parts on an operational level?

1.10 Research questions

The problem and goal are defined, so the research questions can be set. The answers to the following questions should make sure that the research goal can be achieved, and the action problem can be solved.

1) What are the root causes and characteristics of (unplanned) non-availabilities and how can they be identified in advance?

Motivation: To recognize and analyze prospective non-availabilities, it must be clear what circumstances lead to a non-availability. The goal of this research question is to find out what information can be used to reliably categorize any 12NC-Local Warehouse combination on the likelihood of becoming non-available soon.

2) How is alert generation, exception analysis and intervention decision making described in literature?

Motivation: After it is clear what information can be used to predict and prevent probable non-availabilities, I must find a way to transform this into a decision model which can generate alerts and based on that generate an intervention decision. I did a rough scan through available theory, and some literature can be found on the topic of intervention decisions. The goal of this question is to find a framework or method, so I can proceed with the decision model.

3) Which decisions and parameter settings should trigger an alert such that the right number of situations with the highest impact are recognized?

Motivation: The goal of this question is to come up with a good decision model that generates alerts when needed. A parameter is a value with a relation to the variable. An alert might be trigger for example when a variable change above or below the parameter threshold or with an increase more than the parameter. It is important that the generated alerts are reasonable and are worth it to analyze and act upon. Furthermore, a reasonable amount of alerts need to be generated, such that the process is still manageable.

4) How to generate pro-active intervention decisions for the exceptions presented by the Control tower alerts?

Motivation: When alerts indicate on which situations an intervention is desirable, the next step is to research the executable actions to avoid non-availability. The goal of this question is creating a decision model that generate the best intervention for the specific exceptions.

5) What is the impact of an improved Control Tower process?

Motivation: The goal of this question is to quantify the added value of an improved pro-active Control Tower tool and process. This is important input for the conclusion and recommendations of the thesis. It will support implementation decisions of the recommendations, especially when new decision rules, new analysis methods and/or intervention generation techniques are designed but not implemented during the thesis.

1.11 Thesis outline

The remainder of the thesis will be structured as follows: Chapter 2 will focus on describing the current situation of the Control Tower tool and how it generates alerts, how non-availabilities are caused and the performance of the Control Tower tool on Supply Chain Visibility (SCV). Chapter 3 includes a review of the existing literature on alert generation, risk prioritization and operational interventions. It includes also how this literature can or cannot be applied to this project. From there, the issues from the analysis of the current situation and the concepts from literature are formed into a decision model in chapter 4. The numerical results are presented in chapter 5. This chapter will furthermore describe the impact of the new method for ASML. Finally, the conclusions of this project are presented in chapter 6. After the conclusions, a discussion section will dive deeper into the assumptions and limitations of this research. Lastly, recommendations for further research and improvement are presented.

Chapter 2

Current Situation

The "Control Tower"as being used by Customer Supply Chain Management (CSCM) (see section 1.2) at ASML is both an alert generating tool and an exception handling process. This chapter will dive further into the current situation of both the tool and the process. After the description in the first section, the root causes of non-availabilities will be described in the second section. I executed this analysis into root causes to check the non-availability detection of the Control Tower. This results in a description of the so-called "Supply Chain Visibility (SCV)" of the current Control Tower in section three. This includes a description of the input data, what can be calculated based on this information and lastly which issues the Control Tower has in proactively detecting non-availabilities.

2.1 Control Tower

The word "Control Tower"at ASML can refer to the Control Tower business intelligence dashboard and the Control Tower exception handling process. In this report, the word Control Tower tool (or: The tool, depending on the context) is used for the dashboard and the word Control Tower process (or: The process, depending on the context) for the exception handling process.

The tool performs calculations on the input data to generate alerts in four different categories: 1. Demand Sensing 2. Supply Sensing 3. Shortage Within Lead Time 4. Shortage On Lead Time This section will explain these alerts step by step as well as the process to analyze and act upon a triggered alert.

2.1.1 Dashboard

The tool is built using the BI-software package 'Spotfire' and consists of multiple pages (tabs) of visualizations. The main tab shows an overview of all the service parts. The user can filter the overview using graphs by selecting an alert and/or satisfying a certain criticality level (See Appendix A for information on this calculation). When a 12NC is selected, the user can find detailed information in other tabs to reveal more information and analyze the situation that triggered an alert.

The Control Tower tool contains information per 12NC about stock levels, forecast, usage and supply. All this information is aggregated and visualized on the worldwide (network) level. This means the alerts are triggered for issues on a worldwide scale and not for local or regional availability issues. In case of shortages, the automated replenishment application NORA will trigger replenishment to a local warehouse. However NORA does not analyze trends and patterns on local warehouse or regional level, so it can not give a warning on exceptions or threads in the supply chain.

2.1.2 Alerts

This section explains the function of the different alarm types as stated above. The Spotfire Dashboard contains multiple tables joined together. One of the tables contains the alarms in the columns and the unique 12NCs in the rows. An alarm is triggered for an item, when the column of that specific alarm contains a 1. If the alarm is not triggered, the column contains a 0. This section explains how these triggers are being calculated in the current situation.

Demand Sensing

The purpose of this alarm is to indicate when a demand pattern behaves differently then expected with the forecast.

Alert trigger

This alert will be triggered if the worldwide usage of a specific 12NC in a short term and long term period is substantially higher than the forecasted usage in those periods. To make the short term usage pattern more important, a higher weight factor is being used. This rule and the corresponding thresholds have been defined by the developers of the original Control Tower. The threshold values are determined by trial and error during the pilot phase of the alert. This sentence is removed in the version for publication.

The calculation logic can be found in algorithm 1. This is written in pseudocode which does not correspond with the Spotfire syntax.

Decision rule 1 Demand Sensing

if $\sum_{k=t-4}^{t} U_{i,k} \geq 2 * \sum_{k=t-4}^{t} \lambda_{i,k}$ **and** $\sum_{k=t-13}^{t} U_{i,k} \geq 1.5 * \sum_{k=t-13}^{t} \lambda_{i,k}$ **and** $\sum_{k=t-4}^{t} U_{i,k} \geq 3$ **then**
  $DemandSensing_i \leftarrow 1$
else
  $DemandSensing_i \leftarrow 0$
end if

In decision rule 1, $\sum_{k=t-x}^{t} U_{i,k}$ means the Confirmed usage of part i in the period $[t-x, t]$ where t stands for the current week and $t-x$ for x weeks ago. $\sum_{k=t-x}^{t} \lambda_{i,k}$ means the forecasted usage of part i in the period $[t-x, t]$ where t stands for the current week and $t-x$ for x weeks ago. $DemandSensing_i$ is the result of the trigger algorithm for part i.

The factors and period lengths in the algorithm are defined by trial and error. When designing the Control Tower, the goal was to include all cases where forecast review is justified. This sentence is removed in the version for publication.

Limitations

As can be derived from algorithm 1, a demand sensing alert is only triggered when the demand is higher than the expected forecasted usage. When the demand is structurally lower, there is no trigger. This is in correspondence with the goal to maximize material availability. Be reversing the logic, there is also an opportunity for cost reduction. More accurate forecasts will lead to more accurate base stock levels. When these base stock levels can be lowered the base stock level and therefore reducing holding costs.

Analysis

When the alert is triggered for a 12NC, the Control Tower Analyst will analyze the alert and propose an action when a deviation of the norm is detected. These deviations are for instance, a global increase of usage due to reliability issues with the part or local usage increase caused by natural events. The Control Tower Analyst will request feedback from the regional contact persons to find out whether the usage was incidental or if this pattern is expected to be continued. These findings will be communicated with the forecast analyst, who will decide if a forecast increase or local base stock level increase is needed. When it is determined, the usage was incidental, the forecast analyst will prevent a forecast increase by excluding the usage.

Supply Sensing

The purpose of this alert is detecting if there is some change in responsiveness of the supplier(s) of a spare part.

Alert trigger

The trigger for this alert is the difference between the expected goods received quantity (scheduled receipts) and the actual goods received quantity. The expected goods received quantity is based on the standard supplier lead time from the ERP system. In this calculation (see algorithm 2), there is also a distinction between a short term and a long term period both with different weight factors.

Decision rule 2 Supply Sensing

if $\sum_{k=t-4}^{t} SR_{i,k} \geq 3 * \sum_{k=t-4}^{t} AGR_{i,k}$ **and** $\sum_{k=t-40}^{t} SR_{i,k} \geq 1.5 * \sum_{k=t-40}^{t} AGR_{i,k}$
and $\sum_{k=t-4}^{t} SR_{i,k} \geq 4$ **then**
  $SupplySensing_i \leftarrow 1$
else
  $SupplySensing_i \leftarrow 0$
end if

In rule 2 above, $\sum_{k=t-x}^{t} SR_{i,k}$ means the number of items for part i scheduled to be received in the period $[t-x, t]$ where t stands for the current week and $t-x$ for x weeks ago. $\sum_{k=t-x}^{t} AGR_{i,k}$ is the actual quantity of part i received from the supplier in the period in the period $[t-x, t]$ where t stands for the current week and $t-x$ for x weeks ago. $SupplySensing_i$ is the result of the trigger algorithm.

The weight factors and long and short term periods are determined by the developers of the original Control Tower using trial and error during the pilot phase. This sentence is removed in the version for publication.

Limitations

When this alert is triggered, it can indicate either a problem at ASML's supplier or an internal supply issue. Materials arriving from the supplier have to be converted into service parts at ASML's factory. This additional step is included in the supply sensing alert. Unfortunately, there are only a few operational interventions that can be used specifically with supply related issues. One of the options is blocking usage until supply comes in.

A drawback to the decision rule is that when an order is three or four weeks late, no alert will be generated. Furthermore, the alert only holds for fast movers due to the restriction on minimum Scheduled Receipts quantity in 4 weeks. This sentence is removed in the version for publication.

The alert does provide a useful trigger when the internal processes are delaying the supply. An example is a situation where the spare part consists of multiple sub components for which one of the components is not ordered in time. This exception is one of the exceptions that are being detected by the Supply Sensing alert.

Analysis

Not every Supply Sensing alert that is triggered is reasonable to analyze. To come to a reasonable list of 12NCs to analyze, the set of triggered parts is filtered first on the ASML defined worldwide Fill-rate[1] , as shown in equation 2.1, where OH_i stands for the current worldwide stock level of part i, including the central warehouse inventory, incoming supply the pipeline inventory to the local warehouses and the local warehouse inventories. BSL_i stands for the sum of all basestock levels in the central and local warehouses.

$$FillRate_i^{ASML} = \frac{OH_i}{BSL_i} \tag{2.1}$$

Another filter is the responsible planning department. Newly introduced parts (which do not necessarily have to be service parts) are planned by the Product Lifecycle Management (PLM) department, rather than the CSCM department. After a part has had significant usage, the part is transferred to CSCM. Parts outside of the CSCM planning scope are not being analyzed when they have been triggered by the Supply Sensing alert.

Shortage On Lead Time

The Control Tower simulates the future stock levels per week using the forecasted demand and scheduled receipts quantities. This so-called simulation is deterministic based on fixed demand forecast (same for each week) and supplier lead times.

A shortage on local warehouse level is defined as the difference between the Base Stock Level (BSL) of that part at the local warehouse and the current stock level of the part at the local warehouse. A worldwide shortage of a part is defined as the difference between the worldwide target and stock level of a part. The worldwide target is seen as the sum of all BSLs of the local warehouses. The worldwide stock level is the sum of all stock including pipeline stock and incoming supply to the Central Warehouse.

The purpose of this alert is detecting problems or anomalies in the ordering process so that they can be resolved in time and a non-availability can be avoided.

Alert trigger

The Shortage On Lead Time alert is triggered when a shortage is expected at the end of the supplier (new-buy) lead time at the central warehouse of that part. This means a new buy or repair order at the time of an alert should still prevent a shortage, when the supply arrives upon supplier lead time. The reason for this alert is to prevent future shortage caused by e.g. no new buys or repairs are scheduled or when the supplier indicates the ETA is later than the agreed supplier lead time. The root causes for these problems vary from situation to situation and are analyzed by the exception handling analyst.

In rule 3, $\sum_j OH_{i,j,t+L_i^{NB}}$ indicates the future worldwide on-hand stock level (aggregated over all local warehouses j and the Central Warehouse ($j = 0$)) when the (predefined, deterministic)

[1]The ASML defined fill-rate is also calculated for regions and local warehouses. This definition of fill-rate deviates from the definition in literature. This fill-rate fluctuates day by day as inventory depletes with demand and increases with incoming supply or replenishment.

supplier lead time ends. This is calculated by adding the Scheduled Receipts at the Central Warehouse (scheduled between the current week t and when supplier lead time ends at $(t + L_i^{NB})$ to the current on hand inventory and subtracting the predicted usage (monthly forecast as a fixed rate). $\sum_j BSL_j$ represents the worldwide target as sum of all Base Stock Levels in the local warehouses and *excluding* the base stock levels in the central warehouses.

Decision rule 3 Shortage On Lead Time

if $\sum\limits_{j \in J} OH_{i,j,t+L_i^{NB}} < \sum\limits_{j \in J \setminus 0} BSL_j$ **then**
  $ShortageOnLT_i \leftarrow 1$
else
  $ShortageOnLT_i \leftarrow 0$
end if

Limitations

A major limitation of this alert is the projection of stock levels based on deterministic simulation, meaning the stock level depletes each week with the forecasted usage for its week. The stock increases with the scheduled receipts for each week. This means the risk assessment is limited, since the demand of some parts fluctuate heavily and the incoming supply of certain parts is not reliable. This sentence is removed in the version for publication.

Analysis

The Network Oriented Replenishment Application (NORA) should order new parts, or schedule a repair for a defect part, when a shortage occurs. To avoid intervening normal operations, the exception handling analyst does not analyze a triggered 12NC unless the alert reoccurs for some consecutive weeks. This means an exception to the standard operations has occurred rising the risk on a shortage or even a non-availability. The four-time-recurrence filter is decided based on experience.

The analysis consist of a Root Cause Analysis (RCA) to find out why the stock level is (expected to become) lower than the BSL on lead time. This can be caused by lacking (ordering of) supply, local warehouse not returning the defects parts so that they can be repaired, or the BSL of the central warehouse is too low. Currently, the interventions following this alert are triggering new supply of new buy parts or repairing defect parts.

Shortage Within Lead Time

The purpose of this alert is to detect supply expediting opportunities to balance the stock level over time.

Alert trigger

The Shortage Within Lead Time alert is triggered when the target is met right now and when the new buy lead time ends, but a shortage occurs in between. The logic behind this alert is to try to shift the arrival of new stock to an earlier moment (this process is called Re-inning supply).

The alert trigger algorithm is very similar to algorithm 3 for the Shortage On Lead Time. However in rule 4, $\sum_j I_{i,j,t}$ represents the current worldwide on-hand stock level of part i as a sum over all stock I in the central and local warehouses j (where $j = 0$ denotes the central warehouse) including pipeline stock and incoming supply which are included in the stock of the warehouse in which the goods arrive. Furthermore $\exists k \in (t, t + L_i^{NB}) : \sum_j I_{i,j,k} < \sum_{j>1} BSL_j$ states the alert is only triggered when there exists at least one week k in the period between this week t and at supplier lead time L_i^{NB} in which a shortage occurs.

Decision rule 4 Shortage Within Lead Time

if $\sum_{j \in J} OH_{i,j,t+L_i^{NB}} \geq \sum_{j \in J \setminus 0} BSL_j$ **and** $\exists k \in (t, t + L_i^{NB}) : \sum_{j \in J} OH_{i,j,k} < \sum_{j \in J \setminus 0} BSL_j$ **then**
  $ShortageWithinLT_i \leftarrow 1$
else
  $ShortageWithinLT_i \leftarrow 0$
end if

Limitations

This alert type faces the following limitations:

- Re-inning is already executed on an operational basis by another department. Therefore this alert is out of scope for this project.
- When a Shortage On Lead Time issue is resolved by placing a new buy order or repair order, new supply is scheduled to be received on lead time. If there is still a shortage to be expected before the supplier lead time ends, this alert will be triggered as well.

2.1.3 Exception Handling Process

This section is removed in the version for publication.

2.2 Root causes of a non-availability

In this section, the results of analysis into the root causes of a non-availability are described. This analysis is executed to understand which factors are important when trying to detect a non-availability in advance. I started this analysis with an interview with one of the initiators and *owner* of the Control Tower tool and process. From a list of reasons for a part becoming critical, I worked towards a list of root causes of a part being non-available.

2.2.1 Reasons for a part becoming critical

In the first half of 2017, one of the initiators of the Control Tower performed a Root Cause Analysis (RCA) on all parts that received the highest criticality level. This roughly means the part has a low worldwide Fill-rate (the part does not reach the BSL in the local warehouses across the world and central warehouse), and the fill-rate is not expected to reach 100% anytime soon. The top ten exceptions that lead to a high criticality level are listed below.

This section is removed in the version for publication.

The reasons in this list can be assigned to higher level categories as shown in table 2.1. These categories will be used in a further analysis into root causes of non-availabilities

Table 2.1: Categories to which reasons for a high criticality level can be assigned to

This table is removed in the version for publication.

2.2.2 Causes of a part being non-available

The reasons for parts reaching the highest criticality level, are similar to the reasons for a non-availability. This section will diver deeper into the reasons described above.

Demand related causes

This category of causes correspond with the reasons **??**, **??**, and **??** listed in the previous section. Usage peaks (item **??** in the previous list) can have different root causes, namely:

- Natural causes, e.g. an earthquake damaging all shock breakers in a machine
- Replacing parts with productivity or lifetime issues in a group of machines or facing one or multiple DOAs (which overlaps a little with item **??** as well)
- This sentence is removed in the version for publication.

Another usage related reason for a non-availability is usage for analysis (item **??**). This means a part is used for an analysis in a malfunctioning machine for some time, and needs re-qualification and repack afterwards. During the complete operation (including the time the part is in Engineering Stock, the duration of the service event, the shipment towards the supplier, and finally the repair lead time), the part is not available for service. Replenishment will be triggered after the part has been booked to the machine (Confirmed usage) or when the part is sent back to the customer.

This paragraph is removed in the version for publication.

Other usage related categories are a rapid increase of the installed base and initial demand of parts that have no usage data for accurate forecasting.

Quality related causes

The **??**nd most frequently occurring reason for parts reaching the highest criticality level is a quality issue. When a DOA occurs, a second part is needed for the repair, causing a usage peak. A Stock purge means stock is being withdrawn from the local warehouses due to a quality issue. This means a service action cannot take place until a Successor part is available.

Another quality related issue is the shelf life of a part, meaning after a certain expiration date, it cannot be used anymore. This might lead to sudden stock level drops when multiple parts were ordered at the same time.

Supplier related causes

Supplier related causes of a non-availability correspond with reasons **??**, **??**, and **??** of the list in the previous section. Most of the supplier related issues are caused by capacity problems of the supplier. Other causes might be a new supplier or a new Successor are used.

This paragraph is removed in the version for publication.

Stock related causes

This paragraph is removed in the version for publication.

Tactical level related causes

The last category of reasons are related to tactical level related causes like the BSL (item **??**) and forecast (item **??**). The reason why BSL is on the list is the influence of the BSL on the ASML defined Fill-rate. When the BSL increases, the stock level will not immediately be as high, therefore, the Fill-rate will drop. A BSL increase does not lead to non-availabilities, however, a too low BSL means the actual availability of the part is lower than promised to the customers to which the local warehouse is allocated.

2.3 Supply Chain visibility

After analyzing root causes of a non-availability, the ability of the current Control Tower to proactively detect these causes is tested. In literature, the Control Tower capabilities are described as "SCV ", although various definitions of Supply Chain Visibility (SCV) can be found in literature. Each of these definitions has a different perspective on Supply Chain integration and the level on which decisions are made (strategic, tactical or operational). One of the definitions that focuses on Supply Chain integration is as follows:

> "The ability of parts, components or products in transit to be tracked from the manufacturer to their final destination, improving and strengthening the supply chain by making data readily available to all stakeholders, including the customer. "
> (Rouse, 2009)

Two other definitions of SCV are more focused on signaling exceptions in the operational supply chain processes. These correspond better with the vision of ASML with regards to the Control Tower.

> "Software applications that permit monitoring of events across a supply chain. These systems track and trace inventory globally on a line-item level and notify the user of significant deviations from plan. Companies are provided with realistic estimates of when materials will arrive."
> (Vitasak, 2005)

The definition I will refer to in the remainder of this document is as follows.

> "The ability to be alerted to exceptions in supply chain execution (sense), and enable action based on this information (respond). In essence, visibility is a sense and respond system for the supply chain based on what is important in the business."
> (McCrea, 2005)

The ability to generate a signal on exceptions in the supply chain is present in the current Control Tower, although I have to refer to the limitations described in section 2.1.2. The system lacks the ability to show realistic estimates of when materials will arrive. Advanced demand and supply information like proactive maintenance schedules, activity based maintenance schedules and estimated time of arrival of incoming supply is not included in this system. Furthermore, the alarms are all based on information that is aggregated on a worldwide level.

Diving deeper in the sensing ability, I found limitations of the current Control Tower tool categorized per root cause category from the previous section. This is summarized in table 2.2. Of course, the current Control Tower tool, does already show information and generate alerts on most of the non-availability Root Cause categories, however there is still room for improvement on the quality of the alerts, the depth of the information (high worldwide aggregation level instead of locally when possible), the extent to which a decision can be taken automatically, and the frequency in which the data is refreshed (some data is not correct after one week, e.g. stock levels).

Table 2.2: The sensing ability issues of the current Control Tower tool

Root Cause Category	Sensing issues
Demand Related Causes	
	• Natural causes are impossible to detect and predict, although when such an event occurs, no feedback is given to the Control Tower, so that the cause of a usage peak can be shown and communicated. In the Current Situation, the Control Tower analyst receives an alert for a usage peak and needs to ask the region subject to the event to track down the cause of the peak. • This sentence is removed in the version for publication. • This sentence is removed in the version for publication. • This sentence is removed in the version for publication.
Quality Related Causes	
	• Information on shelf life, the duration stock is kept on a SLOC and the SLOC on which an item is stored are not visible in the Control Tower dashboard. These have to be tracked down using other reports or the ERP system. • Information on DOAs and Stock purge events are not visible in the Control Tower.
Supplier Related Causes	
	• No information on the accurate Estimated Time of Arrival (ETA) of a new-buy order is available in the Control Tower, nor is it used to calculate expected shortages. For this the planned receipts or last confirmed date are used (the planned goods arrival date is the supplier lead time in the ERP system, which may deviate from the actual supplier lead time).
Stock Related Causes	
	• Information on shelf life, the duration stock is kept on a SLOC and the SLOC on which an item is stored are not visible in the Control Tower dashboard. These have to be tracked down using other reports or the ERP system. • Information of stock levels is on worldwide level, which leaves the opportunity to redistribute stock between local warehouses and regions by looking at regional or local stock levels and risks.

Tactical level Related Causes	• The forecast accuracy of an item is displayed as the sum of the difference between the actual demand and the forecast in each week over a period of one year. This is already a good indicator of forecasting issues, but the quality of the alerts and the analysis of the alert can be improved, since it does not always make sense to adjust the forecast when the Demand Sensing alert is triggered. • The forecast in the Control Tower, used to calculate expected shortages, is used as a fixed number of parts per week without taking the randomness into account.

2.4 Conclusion

This chapter defined and described how the current Control Tower tool is used in the Exception Handling process. It described how the four alert categories (1. Demand Sensing 2. Supply Sensing 3. Shortage On Lead Time and 4. Shortage Within Lead Time) are calculated and the limitations of these calculations.

From there, I looked at the causes of a non-availability in the categories: 1. Demand related causes, 2. Quality related causes, 3. Supplier related causes, 4. Stock related causes, and 5. Tactical level related causes.

Based on these reasons for a non-availability, I described the SCV abilities of the current Control Tower tool. My conclusions from this analysis are:

1. Due to the limited degree of depth of the decision rules of the existing alerts, too many alerts are generated of which only can be analyzed within the given resources. The way these alerts are selected are non-standardized and prioritization based on criticality level results in firefighting.
2. The current Control Tower alerts are mostly reactive, and when proactive decision making is desired, a new alert type needs to be implemented as well.
3. Due to expanding the scope of the Control Tower Analyst, less alerts can be analyzed and therefore fewer non-availabilites can be avoided by Control Tower proposed interventions.
4. The outcome of the actions proposed by the Control Tower are not logged after the interventions have been executed, making it hard to come to reliable statistics on alert quality an intervention success.
5. The current Control Tower tool has information and alerts on most of the Root Cause categories. Detailed supply chain information on supply and demand side are missed.
6. The stock is aggregated on a worldwide level, which leaves an opportunity to look at local or regional stock levels and risks to proactively redistribute stock and risk levels.

Chapter 3

Literature

In this chapter, I describe the relevant literature into 1. relevant alert triggers for exceptions in the Supply Chain, and 2. relevant interventions after receiving an alert. After every subsection, I compare literature with the practices at ASML. The last section gives a conclusion and describes the next steps for this research.

3.1 Alert generation techniques

A review paper by Topan, Eruguz, Ma, van der Heijden, and Dekker (2018) describes the existing literature on operational interventions that can be proposed by a Service Control Tower. In this paper, two types of intervention triggering methods are described: Time-based and Event-based, which are defined as follows: "Time-based interventions are triggered at fixed points or periods. Event-based interventions are triggered by certain events."Furthermore it is said that: "Reactive interventions are often event-based, and proactive interventions are usually time-based."

This section is structured using the categories of root causes to a non-availability, defined in section 2.2.1. These categories are later used to describe the SCV in section 2.3.

3.1.1 Demand related triggers

This category of triggers should warn for situations where different demand patterns are detected or can be expected. Spare parts demand is the need for replacement of a faulted part. Demand for spare parts can be predicted based on the condition of the components in the installed base. Alerts for operational interventions such as ordering new parts or transshipping parts to other locations can be generated when spare parts are expected to fail soon based on their condition.

In a recent study by Olde Keizer, Teunter, and Veldman (2017), the condition of spare part inventory planning in combination with Condition Based Maintenance (CBM) is discussed. They describe an optimization model and Markov Decision Process model to decide when to order spare parts based on the information on the state of the multiple components in the machine. They conclude that an $(S-1, S)$ policy is not necessarily optimal. "Instead, the maintenance and inventory decisions should both be based on all available information, i.e. on the state of each component, the number of spares on hand, and the size and expected arrival time of each outstanding order."(Olde Keizer et al., 2017) Operational interventions enabled by the condition information are reserving the last spare if both components are close to failure and delaying an order for spares if both components are in good condition.

Other research on the topic of combining CBM with inventory management of multi-component systems is by Van Horenbeek and Pintelon (2015), Wang, Chu, and Mao (2009) , (Wang, Chu, &

Mao, 2008), and Xie and Wang (2008).

Practice at ASML Condition-based maintenance is currently used only limited within ASML. The ProSeLoNext project also has projects focussed on CBM in Service Logistics. Currently demand is predicted using a forecasting technique with an exponentially smoothed average of the usage and the scheduled installed base growth.

The Demand Sensing alert discussed in section 2.1.2 will be triggered when the demand pattern shows abnormal behavior. Since demand patterns anomalies are input for the forecast, this category is discussed further in a later subsection (section 3.1.5).

3.1.2 Quality related triggers

One category of causes for a non-availability is quality related issues. This means the quality of the parts in inventory is physically or administratively insufficient. To prevent that defect or wrong parts are shipped to the customer, a quality system must be in place. Supply Chain Quality Management is a system-based approach to performance improvement that leverages opportunities created by upstream and downstream linkages with suppliers and customers (Foster Jr, 2008). The review paper by Foster includes literature on the combination of supply chain management and total quality management. The concepts discusses are more systematic approaches like Strategic Supply Management, ISO 9000 standards and LEAN six sigma. Although these concepts do help to prevent quality issues, they are not data-driven tools that can be implemented within the Control Tower.

In a more recent review, researchers explain the information architecture that enables real-time quality management and control in the supply chain. (Xu, 2011) Technological innovations that can be used for automated supply chain quality management are service-oriented architecture, Radio Frequency Identification (RFID), multi-agent systems (continuously monitoring of quality factors in supply chains through using autonomous software entities, so-called *agents*, that can interact with its environment and other *agents*), workflow management and the Internet of Things (IOT). From these concepts, RFID and IOT can be useful for monitoring the movement of spare parts between warehouses and customers. An alert can be triggered when the wrong part is shipped or when a part is still in inventory after passing its expiration date.

Practice at ASML Quality related triggers are not incorporated in the Control Tower, however quality of stock is monitored elsewhere in the organization. Tracking the movement of parts across the supply chain is only done administratively in the ERP-system, hence a larger risk on losing or swapping parts.

3.1.3 Supplier related triggers

Issues in the supply of spare parts strongly relate with supplier management. Reviewing and selecting suppliers is mostly considered as a tactical decision, however for completeness, the literature found on supplier performance measurement will be described in this section. One of the processes of measuring supplier performance found in literature is vendor rating. A definition of vendor rating is given by Luzzini, Caniato, and Spina (2014); "the comprehensive and continuous evaluation of supplier performance in terms of products and services delivered."The researchers state that typically, vendor rating focuses on quality, service level, and documentation. Different KPIs are homogenized through a set of weights to obtain a synthetic score for each supplier. In this research, an empirical analysis shows that most of the companies update the analysis every 1-12 months, depending on the criticality of the supplier. Therefore, supplier monitoring triggers time-based interventions.

The performance of the supply chain as a whole is measured in a Supply Chain Performance

Measurement System (SCPMS). A recent literature review on SCPMS byMaestrini, Luzzini, Maccarrone, and Caniato (2017) defines SCPMS as “a set of metrics used to quantify the efficiency and effectiveness of supply chain processes and relationships, spanning multiple organizational functions and multiple firms and enabling SC orchestration.”The focus of SCPMS is communicating the performance of the Supply Chain to all partners in the network. These are first and multi-tier suppliers and first and multi-tier customers, which also relate to SCV.

In a follow up research from Maestrini, Maccarrone, Caniato, and Luzzini (2018), the manner in which performance is communicated with the suppliers is proven to be critical. Furthermore, they found that diverse signals lead to diverse feedback from the supplier, shaping the buyer-supplier relationship management process. Lastly, in a multi-level SCPMS, signals can be send directly to the extended supply chain. Otherwise the signals sent to the first-tier suppliers will likely affect the signals they sent to their own suppliers.

Practice at ASML A department outside CSCM is responsible for supplier relations and performance. However, the current Control Tower does contain an alert for parts with lacking supply. The method is similar to the theory on tracking signals, described in section 3.1.5 The signal not only triggers supplier related interventions, but also usage related interventions. More information on interventions are explained in section 3.2.

3.1.4 Stock related triggers

This category of triggers should trigger an alert when the stock level is insufficient to cover demand. In literature and in practice, two kinds of policies are described continuous (s, S) review models and time-based (T, S) review models. The first policy triggers replenishment when the stock level drops below the reorder-point s. The order quantity is such that the stock level reaches the order-up-to level S. For the (T, S) the on-hand inventory is reviewed every T periods and the replenished to the order-up-to level S.

Other stock related alert trigger operational interventions like lateral transshipments and expediting supply. These concepts are discussed in section 3.2.2, together with their respective triggers.

Practice at ASML ASML uses a $(s-1, s)$ continuous review policy, where replenishment orders are initiated daily by the Network Oriented Replenishment Application (NORA). The application will trigger reactive replenishment in case of a shortage. NORA triggers a new-buy or repair order after the replenishment order is placed. Currently, proactive interventions based on stock levels or non-availability risks are not used at ASML. Only when a local warehouse is out-of-stock while facing customer demand, an alert is given to trigger a reactive transshipment from one local warehouse to another.

Furthermore, ASML uses a criticality ranking to indicate which 12NCs face a high risk with their current worldwide Inventory position. A part is critical when the worldwide inventory position is very low in comparison to the worldwide planned base stock level (see appendix A for more infromation). In those cases, the supplier is notified to speed up supply. During this research, improvements are being developed for this calculation.

3.1.5 Tactical related triggers

With a trigger in this category, the Control Tower user is notified when the tactical parameters do not correspond with the actual Supply Chain operations. At ASML, the tactical parameters like local warehouse base stock levels and the reorder point at the central warehouse are dependend on the forecasted usage. A method to describe the performance of the forecast is the forecast accuracy. A Tracking Signal measures the performance of the forecast periodically and triggers an alert when the forecast is out of control. These two techniques will be discussed next.

Measures for forecast accuracy Commonly used measures to express forecast accuracy as stated by Hemeimat, Al-Qatawneh, Arafeh, and Masoud (2016) and Hyndman and Koehler (2006) are: 1. bias (difference between actual usage and forecasted usage), 2. the Mean Error (ME) (average bias), 3. the Mean Squared Error (MSE) (average of the squared bias), 4. the Mean Absolute Deviation (MAD) (average absolute bias), 5. the Mean Absolute Percentage Error (MAPE) (Average of the ratio bias versus actual usage) or 6. the Mean Absolute Scaled Error (MASE) (where the bias is divided by the average usage change).

The Mean Error estimates the bias of the forecasting method, the Mean Squared Error and the Mean Absolute Deviation are estimators of the variance. The main difference between the MSE and MAD is that the MSE is more sensitive to outliers. (Hemeimat et al., 2016)

Tracking Signals A broadly defined method in literature to bring the forecast under control is using a tracking signal. This is an unbiased measure and can be used for any type of forecasting method. A tracking signal works as an alert trigger for any bias in the forecasting system if the forecasting process is in control. When the tracking signal rises above a control limit (usually 3 standard deviations (Hemeimat et al., 2016)) then the forecasting process is out of control.

In a research by Alstrøm and Madsen (1996) on tracking signal performance, three categories of signals are compared: 1. Standard **Brown** Tracking Signal, 2. High/Low Brown or Trigg and Leach, 3. Standard Trigg and Leach. The Brown signal plots the absolute ratio of cumulative bias devided by the a smoothed average of the Mean Absolute Deviation (MAD) per period (where the parameter alpha determines the smoothing). In the High/Low method, the alpha level is increased every time the Tracking Signal is triggered. The Trigg and Leach signal uses also a smoothed average for the cumulative bias.

The findings of Alstrøm and Madsen are "1. A High/Low method is better for items with positive expectation of significant changes in demand pattern, but the value of the smoothing factors are crucial. 2. The increase in total costs when changes in demand patterns do not turn out to be as significant as expected are small if the Brown Tracking signal is used."

Hoda Sabeti and Jaridi (2016) did research on the performance of the Trigg Tracking Signal by incorporating non-normal noise into the demand distribution. They used simulation to find an estimate for the standard deviation of the Tracking Signal, such that it can be used to determine the 3σ control limits.

Practice at ASML The forecast at ASML is reviewed and calculated monthly. The base stock levels at the central and local warehouses are reviewed periodically or by request. The current Demand Sensing trigger can trigger a review.

3.1.6 Future developments

New developments in the field of Supply Chain Control Towers are focusing on automating and improving the alert generation capabilities. I will discuss two trends in which automation and data analysis are the main focus.

The first trend in industry and Supply Chain management is Industry 4.0. This term refers to the fourth industrial revolution driven by the use of IOT, Cyber-Physical Systems (CPS) and Smart Factories (Hermann, Pentek, & Otto, 2016). This means that machines, products and transportation systems communicate with each other. This is a paradigm shift from centrally controlled production and logistics to decentralized automated control. I refer to Hermann et al. (2016) and Govindan, Cheng, Mishra, and Shukla (2018) for an extensive literature reveiw.

Trappey et al. (2017) develop a roadmap for implementing Industry 4.0 in logistic services based on a review of literature and patents.

Another trend in Supply Chain management is the use of machine learning. At this moment, little is written concerning the integration artificial intelligence, or more specifically machine learning, within the Supply Chain control tower. However, information on this topic can be obtained via businesses and logistics forums.

Machine Learning, introduced by Samuel (1959), is a specific application of Artificial Intelligence (AI) defined as the science of autonomously learning computer systems. The application of IBM's machine learning system Watson for the analysis of exceptions in IBM's supply chain is recommended by (Schwagmeier, 2016). IBM itself writes about the application of its Watson system within logistics in a white paper collaboration with DHL. One of the described use cases, suitable for implementation within the Control Tower, is "a machine learning-based tool to predict air freight transit time delays in order to enable proactive mitigation "(Gesing, Peterson, & Michelsen, 2018). The system can also identify the causes for the delays up to a week in advance to enable better operational schedules.

Machine learning is also applied to other activities in Supply Chain management such as forecasting ((Carbonneau, Laframboise, & Vahidov, 2008), (Sahin, Kizilaslan, & Demirel, 2013)) and inventory management (Jiang & Sheng, 2009).

3.2 Operational interventions

In the review paper on Service Control Towers by Topan et al. (2018), operational intervention are divided into different categories. Four main categories can be dissected from this review, namely: 1. stock allocation, 2. expediting, 3. cannibalization, and 4. capacity planning. These interventions can be executed both stand-alone or in combination with other interventions. In this section, I will only discuss the literature on interventions relevant to an after-sales service Supply Chain comparable to that of ASML. This means that operational interventions regarding to repair shops (capacity allocation, priority scheduling and overtime planning of repair personnel) are out of scope. Furthermore, cannibalization will not be considered as well, since the authority to make this decision is not delegated to the CSCM department.

3.2.1 Stock allocation

In multiple echelon Supply Chain networks, stock allocation rules determine the priority in which downstream echelon warehouses are replenished upon arrival of new buy or repair orders in the upstream echelon. The possible interventions related to stock allocation according to Topan et al. (2018) are:

Proactive stock allocation from upstream As described above, intelligent stock allocation rules have been developed for determining which warehouse receives a part first. I refer to (Topan et al., 2018) for a complete overview of the literature on this topic. My review is not as extensive, since intelligent allocation rules are already implemented and automated in the ASML Supply Chain.

Skipping regular replenishment Pinçe, Frenk, and Dekker (2015) describe a model to cope with obsolescence of spare parts. By skipping regular replenishments, the Base Stock Level (BSL) is slowly reduced. This proactive intervention reduces the risk of having high inventory of obsolete parts. When spare part inventory becomes obsolete because an SLA ended, can be relocated where there is still demand for the spare parts.

(Dynamic) Stock rationing Dynamic stock rationing is applied when service differentiation between customers is used. Dynamic stock rationing optimizes the decision when to satisfy demand or reserve stock for the demand of a customer with in a higher service class. This problem is previously solved by using dynamic programming ((Evans, 1968), (Kaplan, 1969), (Topkis, 1968), and (Teunter & Haneveld, 2008)) and more recently by using simulation ((Fadıloğlu & Bulut, 2010), (Chew, Lee, & Liu, 2013)), although exact evaluation of the problem remains complex. Kranenburg (2006) and Enders, Adan, Scheller-Wolf, and van Houtum (2014) performed research on Stock Rationing in similar conditions to the ASML Supply Chain.

Practice at ASML Obsolescence forms a significant part of the inefficient inventory classification at ASML. Therefore, a single-adjustment or multi-adjustment policy would reduce the cost spent on ordering and holding obsolete parts. Stock allocation interventions are not relevant for the ASML Control Tower, since the automated replenishment tool NORA uses its own allocation rules based on non-availability risks. Service differentiation is also already applied at ASML, since some customers require some spare parts to be reserved at the nearest local warehouse.

3.2.2 Expediting

Expediting supply literally means speeding up the supply processes and receive supply faster. According to Topan et al. (2018), several types of expediting can be distinguished including:

Emergency shipment from upstream In a two echelon system, emergency shipments from the central warehouse (upstream) can be used to support local warehouses (downstream) in case of demand when the warehouse has no on-hand inventory. Howard, Marklund, Tan, and Reijnen (2015) use a combination of expediting and stock allocation. They describe in their paper how to achieve a cost-efficient policy for requesting emergency shipments. They use the ETA of the next replenishment as pipeline information, to decide whether to request an emergency shipment or wait for the replenishment to arrive. The latter case is also called backordering.

For other research on emergency shipments from an upstream location, I refer the reader to: (Gaukler, Özer, & Hausman, 2008) , (Johansen & Thorstenson, 2014) and (Janakiraman, Seshadri, & Sheopuri, 2014) also referred to in the extensive review paper by Topan et al. (2018). Gaukler et al. (2008) analyze the impact of emergency shipments using a standard (Q,R), while Johansen and Thorstenson (2014) and Janakiraman et al. (2014) use a periodic review policy. An extension to Johansen and Thorstenson (2014) is given by Johansen (2018) in which a simplified model is proposed.

Van Wijk, Adan, and van Houtum (2013) describe a model including a Quick Response Warehouse that is able to support a warehouse that run out of stock faster than an emergency shipment from the Central Warehouse. Simple reactive stock allocation rules determine which warehouse receives a part. Reactive lateral transshipments from other local warehouses are not considered. In this paper, results are given for different critical level policies for when to use a quick response action versus an emergency. They describe that an optimal policy would give a cost reduction of 7%. The impact on improved availability and downtime reduction is not discussed.

In a paper by Topan and Van der Heijden (2018), proactive and reactive expediting interventions are simulated for a similar company. These interventions include: lateral transshipments between local warehouses, emergency shipments from the central warehouse, and using pipeline inventory. They conclude that among the proactive shipment types, proactive emergency shipments from central warehouse to one of the local warehouse contribute the most to downtime reduction. For the low priced parts, proactive lateral transshipments also have significant impact on the downtime reduction.

Lateral transshipment "Lateral transshipments within an inventory system are stock movements between locations of the same echelon "(Paterson et al., 2011). For extensive reviews on this subject I refer the reader to Wong, van Houtum, Cattrysse, and Van Oudheusden (2006), Glazebrook, Paterson, Rauscher, and Archibald (2015) In common literature, two types of transshipments are considered: reactive and proactive. Glazebrook et al. (2015) contribute to this by proposing a third hybrid approach in which both proactive transshipments as reactive transshipments are considered.

Banerjee, Burton, and Banerjee (2003) simulate the use of two transshipment policies: lateral transshipments based on availability (TBA) and lateral transshipments for inventory equalization (TIE). In the TBA policy, parts are shipped when the stock level at a local warehouse falls below the transshipment order point. Stock is transshipped from the local warehouse that has the largest excess stock to the local warehouse having the greatest current need. In the TIE policy inventory redistribution occurs no more than once in every review cycle, so that all local warehouses have an equal number of days' supply. the TBA policy appears to be significantly more effective in preventing stockout incidents. Nevertheless, the TIE policy appears to have at least some merit in terms of reducing the severity of shortages. Moreover, under many circumstances, the TIE policy is likely to yield appreciable savings in transshipment costs. The limitation of the TIE policy is the costs that are not directly considered. For this reason, Tiacci and Saetta (2011) modified this model by including transshipment costs in the decision.

A reactive lateral transshipment intervention model allows emergency shipments from other locations in the same echelon. Relevant research on reactive lateral transshipments come from:Evers (2001), Axsäter (2003), Minner, Silver, and Robb (2003), Lee, Jung, and Jeon (2007), Kukreja and Schmidt (2005), and Yang, Dekker, Gabor, and Axsäter (2013).

Proactive and reactive lateral transshipements can also combined with other operational interventions. Hoadley and Heyman (1977) combine proactive lateral transshipments with proacive stock allocation, returning parts to an upstream echelon, disposing parts from the upstream echelon. The limitation of this research is that it is only focused on the costs for a single period. Grahovac and Chakravarty (2001) consider a combination of reactive lateral transshipments with emergency replenishments, emergency lateral transhipments and backordering.

Practice at ASML Expediting of the new-buy supply and the repair process are out of scope for implementation in the Control Tower since this is already initiated by the operational supplier management department if a part receives a critical level.

Currently, lateral transshipments are only used reactive when a local warehouse has insufficient stock to fulfill the service demand. A part is then shipped with emergency (within 2 days) from another local warehouse, preferably within the same region. ASML adopted the Quick Response Warehouse in the form of an 'Emergency Hub warehouse', in case a lateral emergency transshipment from a local warehouse within the same region is not possible. Proactive lateral transshipments are not used within ASML, however with this review, the potential of this intervention is recognized to improve the availability in situations when replenishment from the upstream echelon is limited.

3.3 Conclusion

In this review, I've evaluated the methods described in literature on how to generate alerts and interventions. What became apparent is that the existing alerts in the ASML Control Tower have never been subject of academic research. Furthermore, literature on how to increase Supply Chain Visibility is also scarce. This gap is filled by consultants offering only commercial information on Supply Chain Control Towers.

I have chosen to do a broad literature review on how the sensing and response abilities of the Control Tower can be improved. The sensing abilities of the current Control Tower have not been described extensively in literature. Other methods to signalize issues in the demand, quality, supplier, stock and tactical level categories have been described.

For the stock related issues, visibility can be obtained by segregating the information on stock levels on a regional level. This gives the opportunity to optimize the availability when replenishment

is impossible.
For demand and quality related triggers, further research into the field of industry 4.0 and machine learning is advised. Innovations from industry 4.0 can be used for continuously monitoring the condition of the machines and the movements of inventory. An alert can be generated when new spare parts have to be ordered or when the wrong spare part is picked for a shipment to the customer.
ASML maintains strong strategical relationships with the supplier, that are managed outside the influence of the Control Tower. Therefore, the improvement opportunity for the Supply Sensing alert is limited.
Improvements to alerts that are related to the performance of tactical level decisions such as the forecast or the base stock levels are out of scope for this project.

The table 3.1 below summarizes the proactive operational interventions mentioned in this chapter. The second column explains whether or not these interventions are applicable to the network and Control Tower of ASML.

Table 3.1: Summary of proactive operational interventions

Proactive operational intervention	**Usefulness for ASML**
Proactive stock allocation	A risk based replenishment prioritization algorithm is already implemented via the NORA application.
Proactively skipping replenishment	This intervention is focused on cost optimization instead of availability optimization.
Dynamic stock rationing	This intervention is interesting for ASML to reduce penalty costs when reserving stock for machine down situations, however is not interesting when the goal is to minimize non-availabilities for all demand types.
Proactive emergency shipments from upstream	This intervention is interesting since it increases the availability of the local warehouses with large risk on a non-availability faster. This does require the Central Warehouse to have stock.
Proactive lateral transshipments	This intervention is useful when the Central Warehouse is out of stock and the network is unbalanced. A shipment between regions with a low risk on a non-availability can support regions with a high risk on a non-availability.

Based on the summary above I conclude that suitable interventions for ASML are proactive emergency shipments from the Central Warehouse and Proactive Lateral Transshipments. The Proactive Lateral Transshipments are also relevant since a large portion of the non-availability issues are caused by insufficient supply. Since supplier related issues can not always be solved on an operational level, the Lateral Transshipments can at least make sure that the risk in the regions are minimized and balanced until replenishments are possible again. I have discussed these findings with the company and based on their input and priorities, I chose to implement the Proactive Lateral Transshipments.

Chapter 4

Decision model

In this chapter I will introduce a two-staged alert and intervention generating model to balance and minimize the risks on a non-availability in a network when the upstream Central Warehouse is unable to replenish the shortages in the local warehouses. The first section introduces the reader to the conceptual model of the decision model. The second section formulates how the alarms and interventions are generated. The last section validates the tool and chooses the optimal decision parameter using factorial analysis.

4.1 Conceptual model

The first subsection explains the context of the decision model by positioning it in the Supply Chain Network. Then the goal of the decision model is explained in the second subsection. Next, the structure of the model is explained in relation to the layers of the Control Tower framework. The last two subsections explain the concept of generating alerts and interventions by explaining the key variables and restrictions.

4.1.1 Supply Chain network

The current ASML Supply Chain network as described in section 1.4 is subject to this research of introducing proactive decision making. This network consists of one central warehouse to which spare parts are delivered from the suppliers. Every customer has a Service Level Agreement (SLA) that determines the maximum allowed Downtime Waiting for Parts (DTWP) for every machine. Therefore, around [redacted] warehouses, located in [redacted] different regions service the customer demand for spare parts. Replenishment shipments are triggered according to the $(S-1, S)$ inventory policy. An automated tool analyzes the system daily and schedules replenishments and triggers new buy orders. A prioritization rule determines in which order the local warehouses are replenished in the on-hand stock at the Central Warehouse is insufficient. This rule determines that the first part will be shipped to the local warehouse with the largest expected unplanned non-availabilities in the next month (see appendix B for a precise definition and its limitations). Due to the fast growth of ASML its installed base and increased complexity of the components, it becomes increasingly difficult for suppliers to deliver the requested demand for new spare parts in time making the prioritization rule more important.

Parts are shipped from Central Warehouse to Local Warehouses with one of three different shipment types: Routine, Priority or Emergency, which means the part arrives respectively, within 2 weeks, within 1 week or within 48 hours. These shipment times are used as upper bounds for shipments between two different regions.

When a local warehouse is out-of-stock while a part is requested from the customer, a non-availability occurs. This Non-availability (NAV) is considered as an unplanned NAV when the local warehouse did have a base stock level. The preferred support solution is an Emergency transshipment from a local warehouse that is located in the same region. When this is not possible, because the on-hand stock of other local warehouses in the regions are zero or allocated to another service order, the responsibility is escalated to the worldwide support team. They determine if an Emergency shipment should be initiated from the central warehouse, the emergency hub, located near the majority of the local warehouses, or from a local warehouse located in another region in the network. Proactive operational interventions are not being used in the current Supply Chain network.

4.1.2 Goal

The goal of this research is to introduce a proactive decision model to this Supply Chain network. The Control Tower enables this proactive decision making by giving Supply Chain Visibility (SCV). This means that exceptions in the Supply Chain network are recognized, analyzed and handled by an intervention to improve the availability of spare parts. To enhance the SCV of the current Control Tower used by ASML, I introduce insight in the regional risks on a non-availability. The same definition that is used for the prioritization of replenishment shipments is also used to indicate these regional risks.

The exception that I propose to recognize and solve with this decision model is an imbalanced network. This means that one of the regions in the network has a relatively high risk on a non-availability while another region in the network faces a much lower risk. These imbalanced situations can easily be solved when it is assumed that the central warehouse has infinite supply. This is however not realistic for any company.

In the current situation, there is no process leveling out the imbalanced network (when the Central Warehouse is out of stock and none of the regions have more stock than their BSL), meaning the availability only deteriorates for regions that already have many expected unplanned NAVs. This continues until new supply arrives at the central warehouse, which can take up to a number of months. According to the definition of(McCrea, 2005), to gain this SCV, the decision model must have the ability to be alerted to this exception (sense), and enable action based on this information (respond). The desired response is a risk balancing intervention in the form of a proactive Lateral transshipment between two or more regions (using a Routine or Priority shipment).

I chose to focus on transshipping parts between regions instead of between all local warehouses so that a reduced optimization problem remains (■ regions instead of approximately ■ local warehouses). When a lateral transshipment between two different regions is proposed, a planner should determine from and to which local warehouse the part is sent.

Another reason is that the calculation for expected unplanned non-availabilities in the next month for a local warehouse only includes standard replenishment lead time and does not include the reactive transshipment option that is used within ASML. When a local warehouse is out-of-stock, another local warehouse in the same region can support quickly. This risk pooling effect is included in the calculation of expected unplanned non-availabilities in the next month in a region. Therefore the reduction of non-availability risk by using proactive lateral transshipments is much higher for lateral transshipments between regions than lateral transshipments between local warehouses.

The proposed proactive lateral transshipments should prevent non-availabilities and hence shipment and escalation costs. To break even, approximately 3 priority and 4 routine proactive lateral transshipments can be used for every inter-regional emergency shipment avoided. However for every hour of downtime avoided[1], around 180 priority or 240 routine proactive lateral transshipments can be used.

[1] This is a very rough estimate based on the most common machine type and contract

4.1.3 Structure of the decision model

Figure 4.1[2] shows how I will introduce the sensing and response capabilities to the ASML Control Tower. The figure shows a simplified representation of the Supply Chain network. To sense the undesired exception (imbalanced network) I present an alert generation model which analyzes the expected unplanned NAVs for every unique 12NC per region. When an alert is generated for a specific 12NC, a second response model is used to generate an intervention proposal.

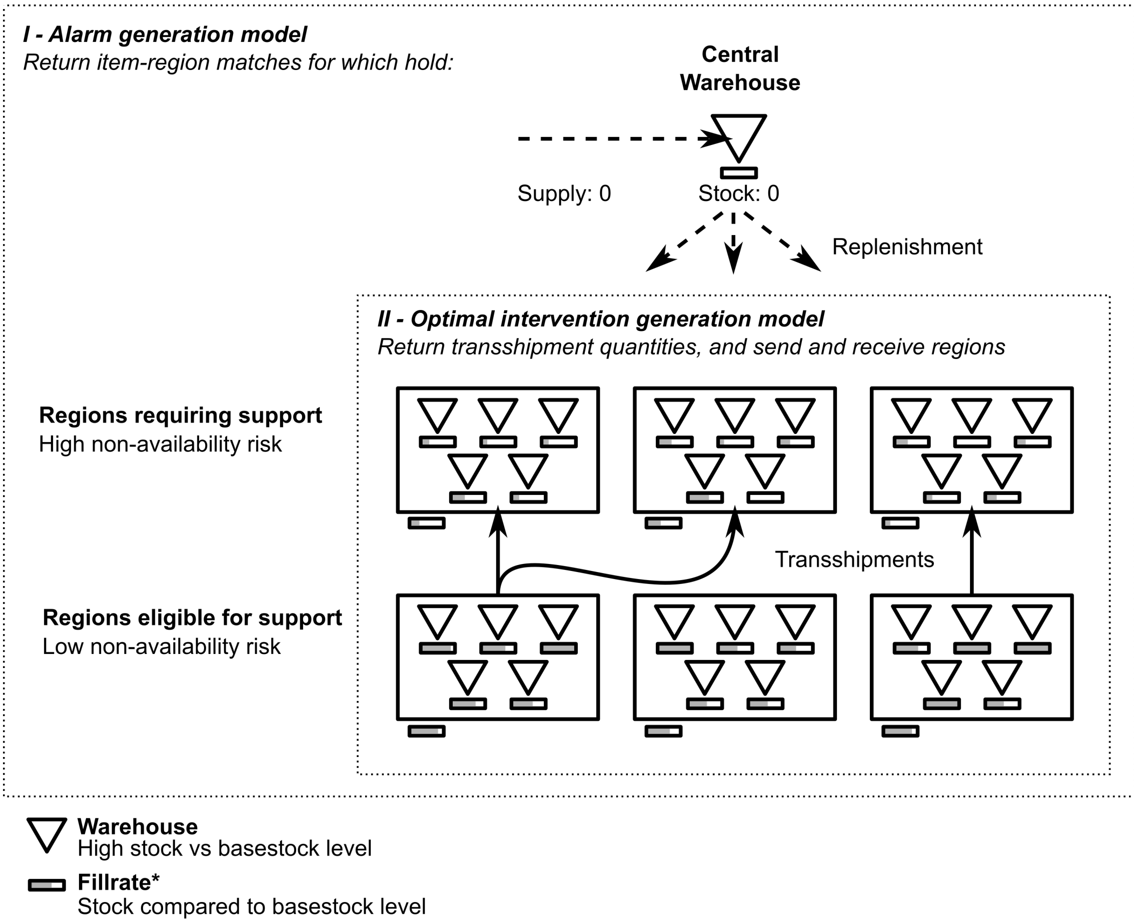

Figure 4.1: Conceptual model Proactive Regional Support Intervention

The two-stage model is based on the layered structure of (Supply Chain) Control Towers, which is described in chapter 1. First, alerts are generated for situations requiring attention from the Control Tower analyst. The alerts are generated and displayed in the Data Application layer. The second step is to generate interventions best suited for the different situations, which is done in the Operational Planning layer. The Control Tower analyst receives decision support from the this intervention generating model that I propose, which optimizes the transshipment quantities between the regions. This concept is shown in the layered structure in figure 4.2.

[2]Please note that the definition of Fill-rate as presented in the figure is ASML its definition of fillrate. Please check the glossary on page ii for more information.

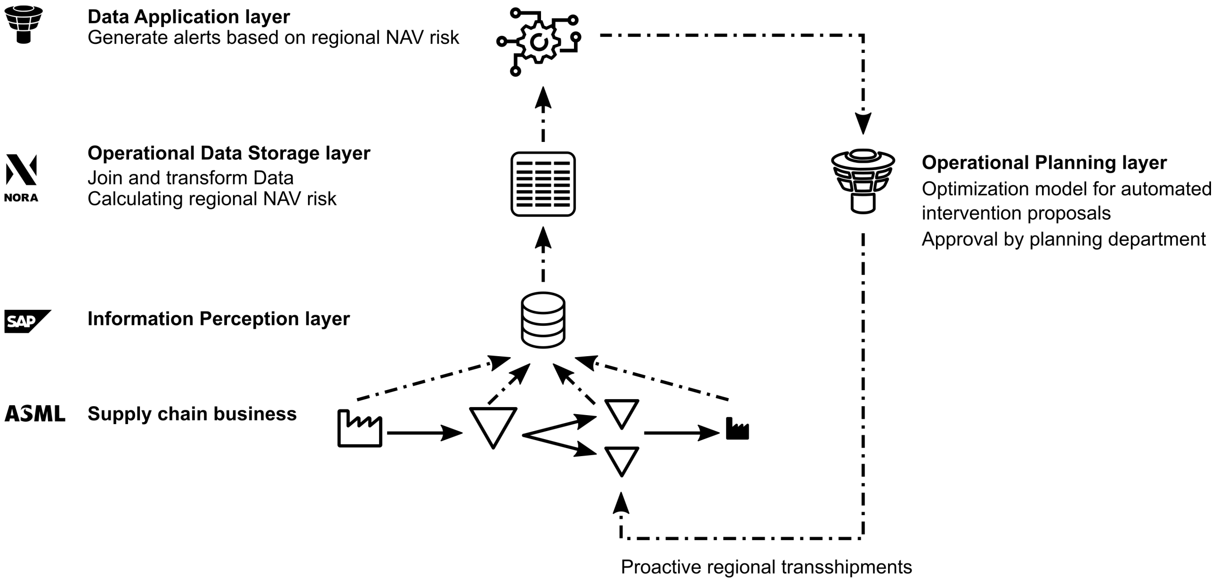

Figure 4.2: Decision model in Control Tower context

Real-time transactions in the Supply Chain are handled through the ERP-system (SAP) in the Information Perception layer. NORA extracts information from SAP and uses this data for a daily analysis on the shortages in the network. A risk-based priority rule decides the replenishment order and priority. This non-availability risk is added to the existing data set in the Operational Data Storage layer. Finally the NORA-data is used in the BI-tool Spotfire located in the Data Application layer. Within this tool, I implemented logic to trigger alerts. This logic is described in the next section 4.2.1. Finally, the alerts are input for the optimization tool in the Operational Planning layer. This optimization tool proposes the best possible intervention based on the model described in section 4.2.2. The information on optimal interventions is given back to the Spotfire tool for an analyst to analyze and approve.

The reason why this two stage structure is chosen, instead of just running this optimization model is the scale of ASML its Supply Chain, with over tens of thousands unique 12NCs with predecessor successor relations and new parts being introduced continuously. The intervention generating model is a linear programming model that proposes intervention to balance the network. Doing this for every 12NC would be very inefficient and can take up to a working day of running time if only a couple of seconds are needed to analyze one 12NC. Therefore the alert generating model makes a pre-selection of 12NCs to optimize.

The proactive operational interventions could be carried out daily. To generate the alerts, data from the automated replenishment application NORA is used, therefore scheduling replenishment shipments and new-buy/repair orders should be carried out first. The interventions do not interfere with the decisions made by NORA. After the interventions have been confirmed and executed, the Inventory Positions are updated and the next day, NORA uses the updated network.

4.1.4 Alert generation

An exception is recognized when a situation of imbalance is detected. Imbalance in this context means, one region has many expected unplanned NAVs, while another region has little expected unplanned NAVs. Therefore the key variable is the expected unplanned non-availabilities. The base stock level at a local warehouse or region is never sufficient to guarantee 100% availability all the time. The number of non-availabilities a region or local warehouses can expect when the stock level matches the bases stock level is called expected planned NAVs. To arrive at the expected unplanned NAVs, the expected planned NAVs are subtracted from the expected NAVs a region or local warehouses faces with the current stock level. The calculation for this variable can be found in appendix B, including its limitations and an alternative approach.

As stated earlier, the alert generating model is used to make a preselection of 12NCs which have an imbalanced network that could be optimized. To recognize an imbalanced network, I have tested different parameters for the decision rules that determine when to trigger an alert. I have tested these parameters and decision rules on business sense (by talking to planners and managers at ASML), significance and correlation. The parameters I've tested were:

The Fill-rate for the receiving and shipping region; The Fill-rate is a widely known and used measure within ASML to indicate availability issues in the network. The CSCM staff expressed their idea to refuse support when the Fill-rate of the supporting region drops below a threshold. This parameter is *not* being used in the decision rules because it is correlated with the expected unplanned non-availabilities in the next month for the region. When the Fill-rate decreases, the risk on a non-availability increases.

The expected unplanned NAVs in the next month; This measure gives an indication of the risk on an unplanned non-availability that a region is facing. This measure can be a threshold for the receiving region, meaning when the risk exceeds this threshold, it is eligible to receive support. Otherwise, it can also be used as a threshold for the supporting region, meaning that the risk must not be beyond the threshold to be eligible to send support. The objective of the model is to reduce the number of unplanned non-availabilities and therefore to reduce the risk on an unplanned non-availability. Therefore, this parameter is used in favor of the Fill-rate.

The on-hand inventory in the Central Warehouse; As described earlier in the conceptual model, when the Central Warehouse has stock on-hand, it should be in transit to a local warehouse to replenish the shortage. Stock in the Central Warehouse should always be sent to a local warehouse, since it can deliver customer demand faster and support another local warehouse in the same region much faster. Therefore, transshipments are not proposed when there is stock in the Central Warehouse.

The scheduled receipts quantity; This is the quantity of parts scheduled to arrive at the Central Warehouse. A normal replenishment from the Central Warehouse approximately takes 14 days. Therefore, when there is supply scheduled within a period of 14 days, the local warehouse is replenished within 1 month. Therefore in reality, when there are scheduled receipts, the risk on an unplanned non-availability in the next month for the regions that are first to be replenished will be lower. Therefore I have decided to set a threshold on the scheduled receipts quantity above which no alert will be generated.
I have expressed this parameter as the percentage of the shortage in the regions needing support. Transshipments are disallowed when more than a certain percentage of the so-called Critical shortage arrives at the Central Warehouse within 14 days.
More control on the allowed scheduled receipts quantity can be obtained by adding a restriction on the quantity scheduled to arrive in the next week as well. I did not experiment with this parameter, since trying every combination with the other parameters would require to much computation time. An experimental design for this parameter is also impossible since it is correlated with the scheduled receipts within two weeks parameter.

The Inventory Position of the supporting region This parameter states the minimum Inventory Position of the shipping region. I have decided to use this parameter in the decision rules, since the Customer Supply Chain Management (CSCM) staff requested me to not allow emptying the supporting region with a transshipment proposal. This means that at least one part has to remain after the intervention. This could lead to a sub-optimal solution in terms of expected unplanned non-availability reduction, however this restriction is desirable for change management purposes. Local Warehouses are in direct contact with service engineers and customers and they are hesitant to "give up"their inventory since to avoid conflict of interests with the customers and engineers.

4.1.5 Intervention generation

The intervention generation model is a decision support tool that presents the Control Tower analyst with ranked intervention proposals. The intervention proposals are generated by using a mathematical optimization problem. The objective function of the model is to minimize the maximum regional expected unplanned non-availabilities in the next month. This means that when a part is transshipped from one region to another, the expected unplanned non-availabilities in the receiving region will decrease while it will increase in the supporting region. When the receiving region was the region with the highest expected unplanned non-availabilities in the next month, the objective function will decrease.

Constraints to the model are: 1. the number of send and received parts must be equal, 2. the supporting region must always keep at least one part, 3. the receiving region can not receive more than its base stock level, 4. in case of excess (more on-hand inventory than the base stock level), the scheduled service orders and sales orders must be supported first, and 5. when the emergency hub is empty, it must be replenished first.

4.2 Model formulation

This section introduces the alert generating and intervention generating model. The first subsection gives the notation and formulation of the alert generating model. The second subsection gives the notation and formulation of the intervention generating model. The last subsection contains 1. a description of the optimal parameter settings, 2. the impact of using different approaches, and 3. the breakdown of the performance into characteristics.

4.2.1 Alert generation

An alert is generated when a situation of imbalance is detected. This occurs when the Inventory Position in one or more regions leads to relatively many expected unplanned non-availabilities and the Inventory Position in one or more other regions result in relatively few expected unplanned non-availabilities. This section treats the decision rules that determine when a region is marked as eligible to receive support and when it is eligible to give support. When a match between a sending and receiving region can be found for a part, an alert is triggered. The next subsection contains the notation of the sets, parameters and variables. Then the calculation for expected unplanned non-availabilities is given. Lastly the decision rules are explained.

Notation

Sets and indexes

$i \in I = \{i_1, \ldots, i_N\}$	The set of all 12NCs ($N \approx 12,000$)
$r \in R = \{r_0, \ldots, r_{15}\}$	The set of all regions
$t \in T$	Set of weeks, where τ is the current week

Parameters

$\rho_{i,r}$	Demand during replenishment lead time
$\lambda_{i,r}$	Monthly forecast for part i in region r
$IP_{a,r}$	The current Inventory Position of the region r for part a
OH_{CWH}	On-hand inventory in the Central Warehouse
$SR_{i,t}$	Scheduled receipts for part i to be received at the Central Warehouse in week t
$\alpha_{\mathbb{E}[NAV]}$	Parameter that indicates the minimum expected unplanned non-availabilities in the next month for the receiving region *before the intervention*
$\alpha_{SR_{II}}$	Parameter that indicates the maximum scheduled receipts quantity in the next two weeks as a percentage of the Critical shortage in the network
$\beta_{\mathbb{E}[NAV]}$	Parameter that indicates the maximum expected unplanned non-availabilities in the next month for the sending region *before* the intervention

Variables

$SN_{i,r}$	*Support Needed*; Binary variable indicating if region r is eligible to receive support for part i
$SP_{i,r}$	*Support Possible*; Binary variable indicating if region r is eligible to give support for part i

A logical parameter to decide if a region is eligible to give support would be the risk on a receiving region *after* the intervention. This parameter is even more able to recognize only reasonable alerts. Due to practical reasons I did not use this parameter since this would require calculating the risks of all 12NCs with one or part received and sent. This is impractical because the Spotfire application is not able to make these calculations, and calculating the risks before importing in Spotfire would require an additional data source in which the calculation is done, hence, even more manual work and additional loading and computation times.

I performed experiments to obtain the best values for the parameters $\alpha_{\mathbb{E}[NAV]}$, $\alpha_{SR_{II}}$ and $\beta_{\mathbb{E}[NAV]}$. In each experiment the values of the parameters are adjusted and each combination yields a different set of 12NCs with an alert and with different regions that are eligible to receive or send a part. Then, I've generated intervention proposals for all the alerts. The performance of the parameter value combination is then determined by the impact of the proposed interventions. This is calculated as the sum of the improvements in maximum regional expected unplanned non-availabilities in the next month over all the intervention proposals. A penalty is subtracted for every alert that does not result in an intervention proposal or when the impact of the proposal is too small. The outcome of the experiments with the resulting proposed parameter values can be found in section 4.3.1

Calculation of the regional Expected Unplanned Non-availabilities in the next month

The NORA algorithm calculates the expected unplanned non-availabilities by first calculating the probability on a non-availability. To do this, the Erlang Loss function (or Erlang B formula) is used, as can be found in equation (4.1). In this equation $c_{i,r}$ represents the stock level of part i in region r. In the Queuing Theory, this probability is also called the "*blocking probability*". This is the probability of not having stock for a part that is requested by the customer. This is a steady state probability, using fixed replenishment lead time and the weighted average monthly forecast as demand.

$$P^{NAV}(c_{i,r}) = \frac{\frac{1}{c_{i,r}!}\rho_{i,r}^{c}}{\sum_{k=0}^{c_i}\frac{1}{k!}\rho_{i,r}^{k}} \qquad \rho_{i,r} = \frac{\lambda_{i,r} * 12}{365} \cdot L_{i,r}^{R} \tag{4.1}$$

From this probability, the expected number of items per month, that cannot be filled from local stock ($\mathbb{E}[NAV]$) can be calculated. NORA does this by multiplying the blocking probability with the monthly forecast of a part, as shown in equation 4.2[3].

$$\mathbb{E}[NAV]_{i,r} = P^{NAV}(c_{i,r}) \cdot \lambda_{i,r} \tag{4.2}$$

The expected unplanned non-availabilities are calculated next by calculating the expected non-availabilities with the current stock level and the targeted stock level.

$$\mathbb{E}[unplannedNAV]_{i,r} = \mathbb{E}[totalNAV]_{i,r} - \mathbb{E}[plannedNAV]_{i,r} \tag{4.3}$$

$$\mathbb{E}[totalNAV]_{i,r} = P^{NAV}(OH_{i,r}) \cdot \lambda_{i,r} \tag{4.4}$$

$$\mathbb{E}[plannedNAV]_{i,r} = P^{NAV}(BSL_{i,r}) \cdot \lambda_{i,r} \tag{4.5}$$

NORA uses the expected unplanned non-availabilities for the allocation instead of the expected non-availabilities per month with the current stock level, since the safety stock levels are based on a so-called "planned risk". If the total risk would be used for allocation, the regions with higher planned risk, can receive a part in favor of a region with higher unplanned risk.

Limitations

A major limitation is that the expectation is calculated by multiplying the probability of facing demand while having no on-hand inventory by the expected demand in a month. A better calculation of expected non-availabilties would be to multiply the probability of having x parts short by x. This suggestion is given by the calculation for expected non-availabilities using Poisson distributed demand is shown in equation 4.6. In this calculation $c_{i,r}$ represents any Regional inventory position of part i.

$$\mathbb{E}BO(c_{i,r}) = \sum_{n=c_{i,r}+1}^{\infty}(n - c_{i,r}) \cdot \frac{\rho_{i,r}^{n} e^{-\rho_{i,r}}}{n!} \tag{4.6}$$

I have compared the two calculations in appendix B. The results of this analysis is that the two calculations yield different expectations, however they are highly correlated (Pearson coefficient of 0.995). This means that the expected (unplanned) backorders of a region are high when the expected unplanned non-availabilities of a region are high and vice versa.

Variables

The logic of the decision rules that determine whether or not an alert will be generated for a 12NC will be discussed in the next section. These rules are programmed in the Business Intelligence (BI)

[3]Statistically this is not the correct way to calculate an expected value, I will adress this issue further in this section, in the discussion in section 6.2 and in appendix B

application Spotfire. This program is able to handle multiple data sets, merge the information and perform transformations and calculations on the data. There are three binary variables that I use in this logic to generate alerts: 1. the *Support Needed* variable ($SN_{i,r}$), 2. the *Support Possible* variable ($SP_{i,r}$) and 3. an *alert trigger* ($Trigger_i$). The dashboard only displays materials for which the alert trigger value equals 1.

Decision rules

Eligible to receive support
A region is eligible to receive support when the expected unplanned non-availabilities in the next month are above the threshold $\alpha_{\mathbb{E}[NAV]}$. Furthermore, the central warehouse does not have any stock on-hand and the stock arriving within 2 weeks is below $alpha_{SR_{II}}$ percent of the Critical shortage in the network. This logic is given in decision rule 5.

Decision rule 5 Eligible to receive support

if $\mathbb{E}[NAV_{i,r}] \geq \alpha_{\mathbb{E}[NAV]}$ **and** $OH_{CWH} = 0$ **and** $\sum_{t=\tau}^{\tau+1} SR_{i,t} \leq \alpha_{SR_{II}}$ **then**
  $SN_{i,r} \leftarrow 1$
else
  $SN_{i,r} \leftarrow 0$
end if

Eligible to give support
A region is eligible to give support when the expected unplanned non-availabilities in the next month for the region are below the threshold $\beta_{\mathbb{E}[NAV]}$. Please note that this is the risk *before* an intervention has taken place. Furthermore, the inventory position of the region must be greater or equal to two. This decision is made to avoid intervention proposals where the supporting region does not have any stock left after the intervention. Although allowing for a region to be empty after an intervention might might lead to a more balanced network, the decision is made by the CSCM staff to restrict this option for change management purposes. In section 4.3 I will investigate the impact of this restriction on the performance. The logic is found in decision rule 6.

Decision rule 6 Eligibile to give support

if $IP_{i,r} \geq 2$ **and** $\mathbb{E}[NAV]_{i,r} \leq \beta_{\mathbb{E}[NAV]}$ **then**
  $SP_{i,r} \leftarrow 1$
else
  $SP_{i,r} \leftarrow 0$
end if

Alert trigger
Finally, when a match is found between a region needing support and a region eligible to give support, an alert is triggered. This logic can be found in decision rule 7.

Decision rule 7 Alert trigger

if $SN_{i,r} = 1$ **and** $\sum_{\hat{r} \in R \setminus r} SP_{\hat{r}} \geq 1$ **then**
  $Trigger_i \leftarrow 1$
else
  $Trigger_i \leftarrow 0$
end if

4.2.2 Intervention generation

The intervention generating model is formulated as a linear optimization model. This section explains the model step-wise by first introducing the mathematical notation of the sets, parameters and variables. The next subsection explains which decision variables are used. Then, the goal function of the optimization model is introduced. Lastly the complete linear programming model is given and the solving method is explained.

Notation

Sets and indexes

$a \in A = \{a_1, \ldots, a_N\}$	The set of all 12NCs with an alert (N number of alerts)
$r \in R = \{r_0, \ldots, r_{15}\}$	The set of all regions
$c \in \mathbb{Z} \cap [-k, k]$	The set of possible Inventory Position changes.

Parameters

$IP_{a,r}$	The current Inventory Position of the region r for part a
$SP_{a,r}$	Binary variable in the alert generating model. Indicates if region r is eligible to *give* support for part a
$SN_{a,r}$	Binary variable in the alert generating model. Indicates if region r is eligible to *receive* support for part a

Variables

$IP'_{a,r}$	The Inventory Position of a region after an intervention
$x_{a,r,c}$	The binary decision variable determining the Inventory Position change of a region after the intervention (see next subsection)

For the range of possible Inventory Position changes, I chose the value $k = 10$, since the size of the Inventory Position of a region is in the same order. When looking at all 12NCs for spare parts, a region has a Base Stock Level (BSL) higher than 10 in only 0.28% of the cases. This, in combination with the goal to balance the risk between regions and not reallocating complete stock levels, makes that this range not a restriction and large enough. In 131 unique cases, there was only one intervention proposal to ship 10 parts from a region to two other regions. Transshipping 11 parts would not be preferred in this case. The model is highly flexible such that the number of possible delta's can be increased or decreased when desired by the Control Tower analyst.

Please note that the Inventory Position of a region $IP_{a,r}$ already contains the quantity replenished as a result of the scheduled receipts. By using the NORA prioritization list, I allocate the scheduled receipts in advance, since they are scheduled to be received before the end of the next month. Therefore, the risk on a non-availability in the next month in that region will be lower to begin with.

Decision variables

$x_{a,r,c}$ decides which delta the current Inventory Position should be adjusted. The binary variable equals 1 for the selected Inventory Position change $c \in C$ for part a in region r. The expected unplanned non-availabilities in the next month are calculated in advance for every possible Inventory Position change and multiplied with the corresponding binary variable, so that only the expected unplanned non-availabilities remain that are a result of the proposed Inventory Position change. I chose this approach because the calculation of expected unplanned non-availabilities in the next month is non-linear. When the decision variables are input to this calculation and the output is used as an objective value, a non-linear solving approach should be used.

Goal function

I've distinguished two nuances for optimally transshipping spare parts between regions. The first approach is to minimize the maximum risk resulting in a balanced network, while the second approach is to minimize the sum of the regional risks. In the first approach, the network is balanced such that all regions have approximately equal risk on facing a non-availability. Apart from an overall non-availability improvement, this options also attributes to reducing the emotional reactions of customers and regions on a non-availability when other regions could have easily supported proactively. The second approach focuses on reducing the total number of non-availabilities in the network. Intuitively this method would be picked over the first formulation, however with this objective, the network after an intervention can still be imbalanced.

The first approach to the goal function; minimizing the maximum regional expected unplanned non-availabilities for the next month, can be formulated as:

$$\text{minimize} \quad z_a \tag{4.7}$$

$$\text{subject to} \quad z_a \geq \sum_C x_{a,r,c} \cdot \mathbb{E}[NAV]_{a,r,c} \quad \forall r \in R \tag{4.8}$$

In equation (4.7), the value of z_a is minimized. z_a is defined in equation (4.8) as the maximum regional expected unplanned non-availabilities. The right hand side of the constraint in equation (4.8) expresses regional expected unplanned non-availabilities for the next month for part a in region r after the intervention, where c equals the proposed change to the existing Inventory Position of the region. The left-hand-side contains the objective value z_a. This value must be greater or equal to the expected unplanned non-availabilities in the next month for every region, meaning that it will be equal to the maximum regional risk.

The second approach to the goal function; minimizing the sum of regional expected unplanned non-availabilities for the next month, can be formulated as in equation (4.9).

$$\text{minimize} \quad \sum_R \sum_C x_{a,r,c} \cdot \mathbb{E}[NAV]_{a,r,c} \tag{4.9}$$

In equation (4.9) the sum of the regional (all regions in the set R) expected unplanned non-availabilities for the next month after intervention (sum over all intervention possibilities in set C) are minimized.

In the objective function, the calculation for the expected unplanned non-availabilities in the next month is given with an additional index c. The adjusted calculation is given in equation (4.10). Before the intervention, the expected unplanned non-availabilities are calculated with the current Inventory Position of a region, with the Inventory Postition change $c = 0$. The intervention generating model proposes an Inventory Postition change $c \in \mathbb{Z} \cap [-k, k]$.

$$\mathbb{E}\left[unplannedNAV\right]_{a,r,c} = \mathbb{E}\left[totalNAV\right]_{a,r,c} - \mathbb{E}\left[plannedNAV\right]_{a,r} \tag{4.10}$$

$$\mathbb{E}\left[totalNAV\right]_{a,r,c} = P^{NAV}\left(IP_{a,r} + c_{a,r,c}\right) \cdot \lambda_{a,r} \tag{4.11}$$

$$\mathbb{E}\left[plannedNAV\right]_{a,r} = P^{NAV}\left(BSL_{a,r}\right) \cdot \lambda_{a,r} \tag{4.12}$$

Constraints

Inventory position of a region after an intervention

The first constraint calculates the Inventory Position of a region r after an intervention has been executed $IP'_{a,r}$. It is defined in equation (4.13) as the current regional Inventory Position plus the change c proposed by the model. c is an integer number between -10 and 10. The binary decision variable is only equal to 1 for one intervention option (when a change of the inventory position of a region is not desired, $x_{a,r,0}$ equals 1).

$$IP'_{a,r} = IP_{a,r} + \sum_{C} c \cdot x_{a,r,c} \qquad \forall r \in R \tag{4.13}$$

The quantity of parts sent must equal the quantity of parts received
The second constraint restricts that when a part is received at one region (the Inventory Position changes with positive c), another region should have sent that part (the Inventory Position changes in this region with negative c). Equation (4.14) states that the sum of all regional Inventory Position changes should equal zero (increases and decreases are equal).

$$\sum_{R} \sum_{C} c \cdot x_{a,r,c} = 0 \tag{4.14}$$

A decision on the Inventory Position after the intervention should be made
$x_{a,r,c}$ is a binary value that equals 1 for the proposed Inventory Position change. When it is not desirable for one region to receive or send a part, the decision variable $x_{a,r,0}$ equals 1. Only one intervention can be made (a region cannot have an increase and decrease at the same time). Furthermore a decision must be made as well. Equation (4.15) states that the sum of all decision variables for a region should equal 1.

$$\sum_{C} x_{a,r,c} = 1 \qquad \forall r \in R \tag{4.15}$$

The Inventory Position can only decrease in a region when support is possible
In section 4.2.1, I explained that the alert generation model decides when a region is eligible to send parts to other regions for support. The binary variable $SP_{a,r}$ (*Support Possible*) only equals 1 when the region r can support part a to other regions, otherwise it equals 0. This variable is used in equation (4.16). Here, the sum of the decision variables indicating that the region has to support another region (Inventory Position change $c < 0$) can only be 1 when $SP_{a,r}$ equals 1.

$$\sum_{c=-k}^{-1} x_{a,r,c} \leq SP_{a,r} \qquad \forall r \in R \tag{4.16}$$

The Inventory Position can only increase in a region when support is needed
In section 4.2.1, I explained that the alert generation model decides when a region is eligible to receive parts from other regions to reduce the risk on a unplanned non-availability in that region. The binary variale $SN_{a,r}$ (*Support Needed*) only equals 1 when the region r is eligible to receive part a from other regions, otherwise it equals 0. This variable is used in equation (4.17). Here, the sum of the decision variables indicating that the region receives a part (Inventory Position change $c > 0$) can only be 1 when $SN_{a,r}$ equals 1.

$$\sum_{c=1}^{k} x_{a,r,c} \leq SN_{a,r} \qquad \forall r \in R \tag{4.17}$$

At least one part remains in the supporting region
As discussed in section 4.2.1, a region can only support when at least 1 part remains in the region after the intervention. This constraint does restrict an optimal balance when a region with low risk has to remain its part at the expense of a region with higher risk. This constraint is however introduced for the purpose of change management during the implementation. This ensures the supporting regions that at all times one part remains. During the validation in section 4.3, I challenge this constraint in both the alert generating model and intervention generating model.

In equation (4.18) the Inventory Position after the intervention must be greater or equal to the binary variable $SP_{a,r}$. This means that when a region is eligible to support, the Inventory Position after the intervention should at least be 1.

$$IP'_{a,r} \geq SP_{a,r} \qquad \forall r \in R \tag{4.18}$$

Sign restriction
The decision variable $x_{a,r,c}$ is binary as restricted in equation (4.19)

$$x_{a,r,c} \in \mathbb{B} \qquad \forall r \in R, \forall c \in C \tag{4.19}$$

Linear model

The next two sets of equations summarize the linear programming models. The first model describes the linear programming with the goal function that minimizes the maximum regional expected unplanned non-availabilities in the next month. The second model describes the options that minimizes the sum of all regional expected unplanned non-availabilities in the next month.

Option 1: Minimize the maximum regional risk

$$
\begin{array}{llcll}
\text{minimize} & z_a & & & \\
\text{subject to} & z_a & \geq & \sum_C x_{a,r,c} \cdot \mathbb{E}[NAV]_{a,r,c} & \forall r \in R \\
& IP'_{a,r} & = & IP_{a,r} + \sum_C c \cdot x_{a,r,c} & \forall r \in R \\
& \sum_R \sum_C c \cdot x_{a,r,c} & = & 0 & \\
& \sum_C x_{a,r,c} & = & 1 & \forall r \in R \\
& \sum_{c=-k}^{-1} x_{a,r,c} & \leq & SP_{a,r} & \forall r \in R \\
& \sum_{c=1}^{k} x_{a,r,c} & \leq & SN_{a,r} & \forall r \in R \\
& IP'_{a,r} & \geq & SP_{a,r} & \forall r \in R \\
& x_{a,r,c} & \in & \mathbb{B} & \forall r \in R, \forall c \in C
\end{array}
$$

Equation set 1: Option 1: Optimization model for optimal transshipment interventions

Option 2: Minimize the total regional risk

$$
\begin{array}{llll}
\text{minimize} & \sum_R \sum_C x_{a,r,c} \cdot \mathbb{E}[NAV]_{a,r,c} & & \\
& IP'_{a,r} = & IP_{a,r} + \sum_C c \cdot x_{a,r,c} & \forall r \in R \\
& \sum_R \sum_C c \cdot x_{a,r,c} = & 0 & \\
& \sum_C x_{a,r,c} = & 1 & \forall r \in R \\
& \sum_{c=-k}^{-1} x_{a,r,c} \leq & SP_{a,r} & \forall r \in R \\
& \sum_{c=1}^{k} x_{a,r,c} \leq & SN_{a,r} & \forall r \in R \\
& IP'_{a,r} \geq & SP_{a,r} & \forall r \in R \\
& x_{a,r,c} \in & \mathbb{B} & \forall r \in R, \forall c \in C
\end{array}
$$

Equation set 2: Option 2: Optimization model for optimal transshipment interventions

Choice for objective function As explained earlier in section 4.2.2, there is a subtle difference between the two objective functions. The goal function that minimizes the maximum regional risk on a non-availability has the objective to balance the network as good as possible, so that the risk on a non-availability is equally distributed between the regions. The second objective function minimizes the overall regional risk in the network. The second options is attractive on a cost perspective, however is possible to leave a larger risk on facing a non-availability in one region, while another region (for which the risk on a non-availability reduces more with one additional part) still has sufficient stock. The planners in the CSCM department experience a lot of emotions when such a situation occurs. Therefore the first problem formulation is preferred by ASML.

I have compared the performance of the two formulations, which can be found in section 4.3.

Optimization solver

To solve the linear programming model, I use an Excel plugin called OpenSolver. This tool is able to handle larger problem instances with more variables and constraints in comparison to the built-in Excel Solver. The required information for the linear model is exported from the Spotfire dashboard and can be copied into the Excel Worksheet. By using a VBA macro, the interventions are generated automatically for all the 12NCs with an alert. The generated interventions can than be copied into an input file for the Spotfire dashboard to present all the relevant information to the Control Tower analyst.

Solving the intervention generating model for one 12NC with an alert takes on average about 2 seconds. This is includes the automated steps that copy the results into an input file that can be imported into the dashboard.

4.3 Validation and numerical experiments

This section describes the outcome of the validation. I have validated the model by doing experiments and studying the performance of the tool under different conditions. Furthermore, I have interviewed multiple planners from the FMA and Service Management departments. FMA is responsible for the operational planning of spare parts in the network. The Service Management department is responsible for the tactical planning of base stock levels and the exception handling using the Control Tower. The first subsection describes how the parameter settings for the intervention generating model are determined. Then a sensitivity analysis is performed to see the impact of decisions that I have made. Finally, I break the set of alerts down into different characteristics to see if some parts or regions need different parameters.

4.3.1 Parameter settings

The goal of the alert generating model is to recognize all the 12NCs that have an imbalanced network. However, it is undesirable that situations are being flagged as requiring an intervention, while it would actually not be beneficial to execute a transshipment (either due to non-availability reasons or to economic reasons). It would be inefficient and undesirable to analyze these alerts and generate interventions for them in the first place. Therefore, the decision parameters $\alpha_{\mathbb{E}[NAV]}$, $\alpha_{SR_{II}}$ and $\beta_{\mathbb{E}[NAV]}$ need to be optimized in such a way that the network is balanced optimally for all the relevant 12NCs and that there are no alerts for 12NCs for which a proactive lateral transshipment is impossible or inefficient. Using this logic I have derived a formula to express the performance of a combination of parameter values. The performance of the parameter set is equal to the sum of the objective value improvements minus a penalty for alerts without an intervention and a smaller penalty for alerts with an intervention with very little impact. The penalties are determined such that the performance of a set decreases when the additional alerts do not generate interventions with substantial impact. I chose a high penalty for alerts that lead to no intervention and a smaller penalty for alerts that lead to an intervention proposal with small impact. The reason for this is that it is undesirable to run the intervention generating model without result. Interventions with small impact are undesired, however it is the choice of the planner when to execute and intervention proposal or when not to take the costs and risks. During the validation it became clear that the a reduction of the maximum risk lower than 0.1 unplanned non-availabilities in the next month is a good cut-off value.

$$\delta_a = \text{Objective value before intervention for part } a - \text{Objective value after intervention for part } a \tag{4.20}$$

$$\text{Performance} = \sum_{a \in A} \delta_a - 0.75 * \text{Count of alerts for which no intervention is proposed} - 0.25 * \text{Count of alerts for which } \delta_a < 0.1 \tag{4.21}$$

To find the optimal parameter setting I performed 143 experiments, each with a different parameter setting, varying the parameters between the values found in table 4.1. I have plotted the surface of the alert trigger performance when two of the three parameters are changed. All test were executed for data from Wednesday 15^{th} of august, which is exactly in the middle of the week and month. The objective function that minimizes the maximum regional expected unplanned non-availabilities is chosen since this option is preferred by ASML.

Table 4.1: Range of values for the decision parameters

Parameter	Range	Step size
$\alpha_{\mathbb{E}[NAV]}$	0.1 - 0.7	0.1
$\alpha_{SR_{II}}$	0% - 100%	10%
$\beta_{\mathbb{E}[NAV]}$	0.1 - 0.7	0.1

Expected unplanned non-availabilities for the receiving region vs. scheduled receipts
First, I varied the values of the expected unplanned non-availabilities in the next month for the receiving region and the percentage scheduled receipts within two weeks at the Central Warehouse. For this series of experiments, the maximum level of expected unplanned non-availabilities for the shipping region is kept at 0.5. I have plotted the resulting performance of the set of decision parameters in a surface plot. The goal is to find the set of parameter values that results in the highest performance.

The results of the first series of experiments can be found in figure 4.3 where the performance of the parameter set is plotted as a surface. The maximum performance can be found for the values: $\alpha_{\mathbb{E}[NAV]} = 0.3$ and $\alpha_{SR_{II}} = 40\%$ where $\beta_{\mathbb{E}[NAV]} = 0.5$ for all experiments.

The relation that becomes clear from this series of experiments is that when the allowed scheduled receipts percentage is increased, more alerts will be generated for which no intervention is proposed. Hence, the performance decreases. This is caused by the fact that the alerts are triggered based on the expected unplanned non-availabilities in the next month of a region, without incorporating the replenishment quantity that will most likely arrive as well in this month. The intervention generating model does incorporate this supply and adjusts (read: decreases) the expected unplanned non-availabilities in these regions. This means that the impact of an intervention is smaller or interventions are not even proposed.

The same relation holds for a very low threshold on the expected unplanned non-availabilities in the next month for the receiving region. This makes sense because when regions with relatively few expected unplanned non-availabilities in the next month are allowed to receive, the reduction of the already low risk will be very small. Lastly, when the threshold on the expected unplanned non-availabilities is increased, the number of useful alerts decreases.

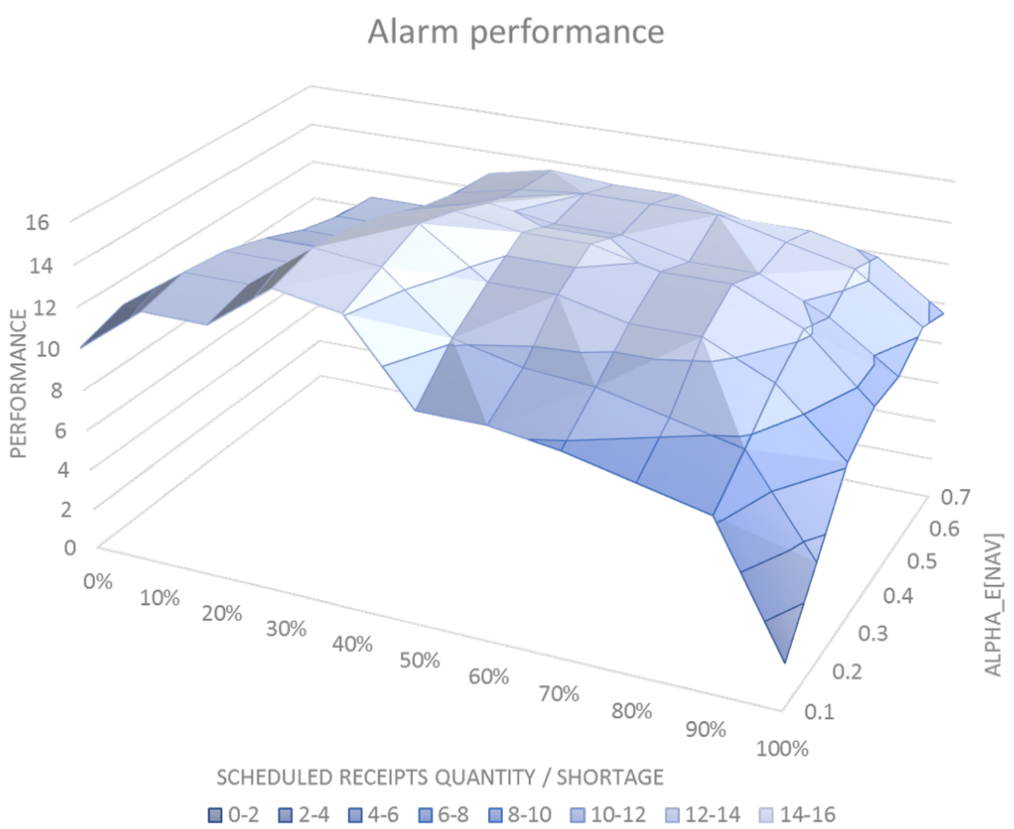

Figure 4.3: Surface plot of alert trigger performance

Expected unplanned non-availabilities for the supporting region vs. scheduled receipts
To optimize the value of the maximum level of expected unplanned non-availabilities at the supporting region, I've varied the values of $\beta_{\mathbb{E}[NAV]}$ and $\alpha_{SR_{II}}$ while keeping $\alpha_{\mathbb{E}[NAV]}$ at 0.3, the optimal value found in the previous series of experiments.

The results of the second series of experiments can be found in figure 4.4. The surface shows no effect of the parameter on changes in expected unplanned non-availabilities in the next month of the shipping region. The main effect of the parameter on the percentage scheduled receipts of the Critical shortage is similar to the effect shown in the first series of experiments. Because of the negligible main effect of the parameter $\beta_{\mathbb{E}[NAV]}$ I decided not to perform 35 additional experiments to plot the surface of the performance with $\alpha_{\mathbb{E}[NAV]}$ and $\beta_{\mathbb{E}[NAV]}$ variables.

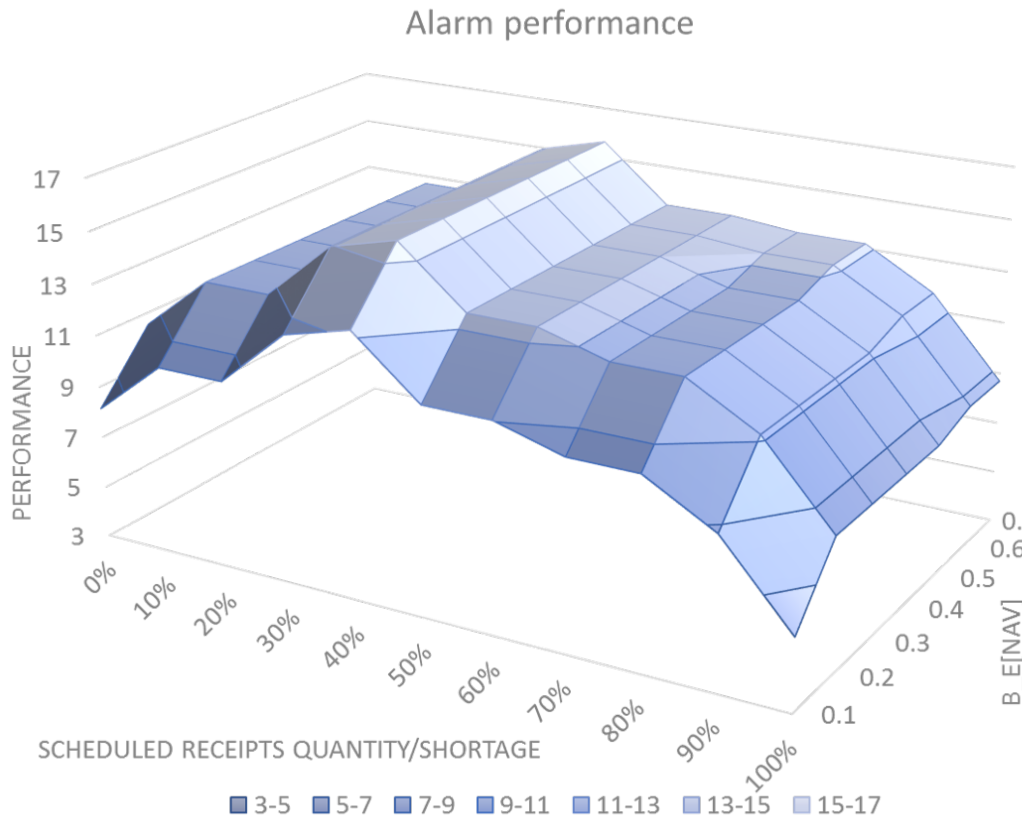

Figure 4.4: Surface plot of alert trigger performance

Choice of parameters In the first series of experiments I have found the best performance of the set of alerts with the parameters set to $\alpha_{\mathbb{E}[NAV]} = 0.3$ and $\alpha_{SR_{II}} = 40\%$ where $\beta_{\mathbb{E}[NAV]} = 0.5$ for all experiments. In the second series of experiments, I have found that $\beta_{\mathbb{E}[NAV]}$ does not have influence on the performance of the set of alerts when $\beta_{\mathbb{E}[NAV]} > 0.1$. Furthermore, it does not make sense that the $\beta_{\mathbb{E}[NAV]}$ is higher than the $\alpha_{\mathbb{E}[NAV]}$, otherwise it is possible that alerts are generated in situations where it can be possible for a region to be eligible to both send and receive support for a part. Therefore, I propose the parameter settings to be used in the alert generating model, as shown in 4.2.

Table 4.2: Proposed parameter settings

Parameter	Description	Optimal value
$\alpha_{\mathbb{E}[NAV]}$	Minimum expected unplanned non-availabilities in the next month for the receiving region.	0.30
$\alpha_{SR_{II}}$	Maximum scheduled receipts quantity in the next two weeks as a percentage of the Critical shortage in the network.	40%
$\beta_{\mathbb{E}[NAV]}$	Maximum expected unplanned non-availabilities in the next month for the sending region.	0.30

4.3.2 Sensitivity analysis

In this subsection I describe 1. what the impact is of using a different objective function, 2. what the impact is of allowing a region to be empty after an intervention, and 3. what the impact is of monthly generating, analyzing and executing interventions instead of weekly.

Objective function

In the model formulation in section 4.2.2 , I introduced two different possible objective functions: 1. Minimizing the maximum regional expected unplanned non-availabilities in the next month so that the risk is balanced between regions, and 2. Minimizing the sum of regional expected unplanned non-availabilities in the next month so that the complete network face as little non-availabilities as possible.

To see what the differences between the two formulated objective functions are, I've performed an analysis on the earlier used test set. When using the optimal parameter settings found earlier, the difference in performance between the two formulations is small. While the first formulation (minimizing the maximum) achieves a bigger *reduction* in the maximum regional expected unplanned non-availabilities in the next month, the second formulation (minimizing the sum of regional expected unplanned non-availabilities in the next month) achieves a higher *reduction* in the overall risk.

The differences are shown in figure 4.5. In this chart you can see the performance of the first formulation (Min Max) and the second formulation (Min Sum) on *reducing* the maximum regional risk and the overall regional risk. The difference fairly small. The efficiency of both alerts is equal, however, the second objective function proposed to transship two parts more. The last column shows the number of alerts for which at least one transshipment is proposed (one intervention can consist of multiple transshipments). This outcome suggests that the impact of using the second formulation is small and therefore the preferred first objective function formulation is chosen to implement.

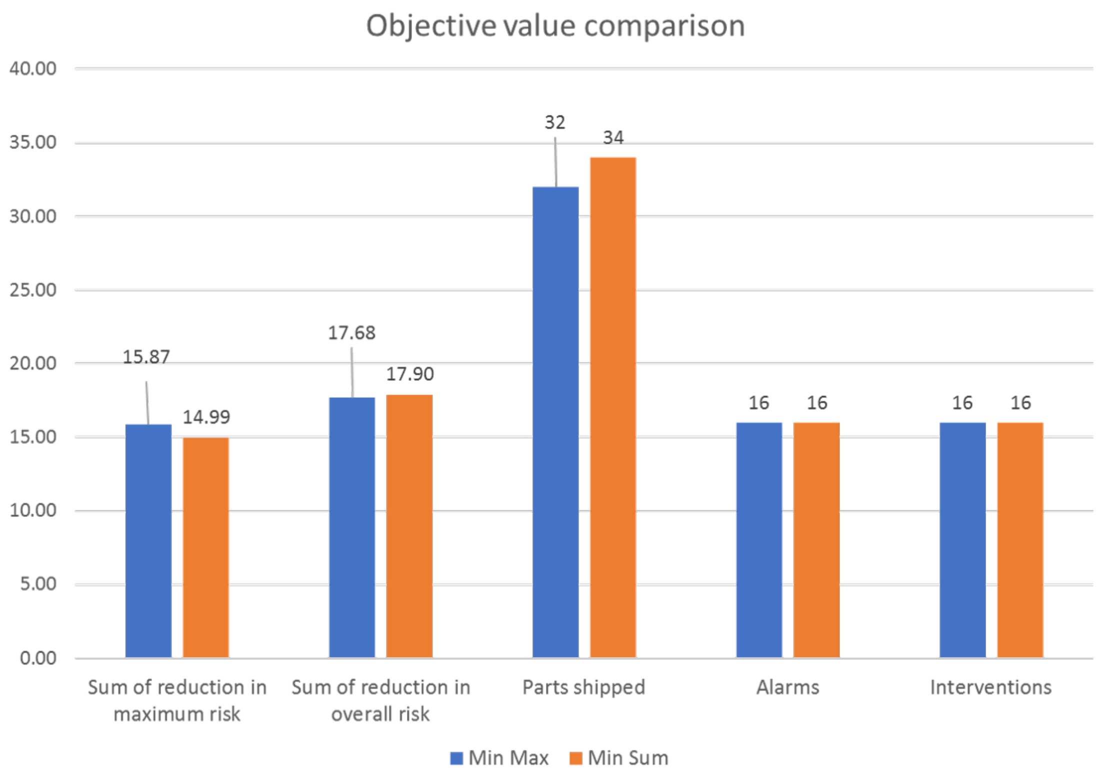

Figure 4.5: Comparison between the two objective function options

Allowing empty regions after an intervention

As described earlier, ASML prefers to leave at least one part in the supporting region as this guarantee supports the change management required to make regions support each other. An analysis on the impact of this restriction is shown in figure 4.6. This analysis is performed on the same data set from august 15, 2018. First, in the alert generation decision rules, the restriction on the Regional inventory position of the supporting region is decreased from at least 2 to at least 1 part. This yields alerts for 4 new 12NCs (from 16 to 20 12NCs with an alert). Dropping the restriction results in an additional 30% improvement of the objective value and an additional 60% improvement of the overall expected unplanned non-availabilities in the next month. Another finding from this analysis is the increase in number of parts proposed to be transshipped. This increases from 32 to 62 parts. The number of shipments increase from 46 to 61. This makes sense since more parts are allowed to be shipped. The number of shipments increase because 1. regions with a BSL of 1 part are now also allowed to support, and 2. the additional part that can be supported can be shipped to another region.

My recommendation for ASML is to consider to drop this constraint, so that a more balanced network can be achieved. Especially slow moving parts will benefit since they often have a regional BSL of only 1 part.

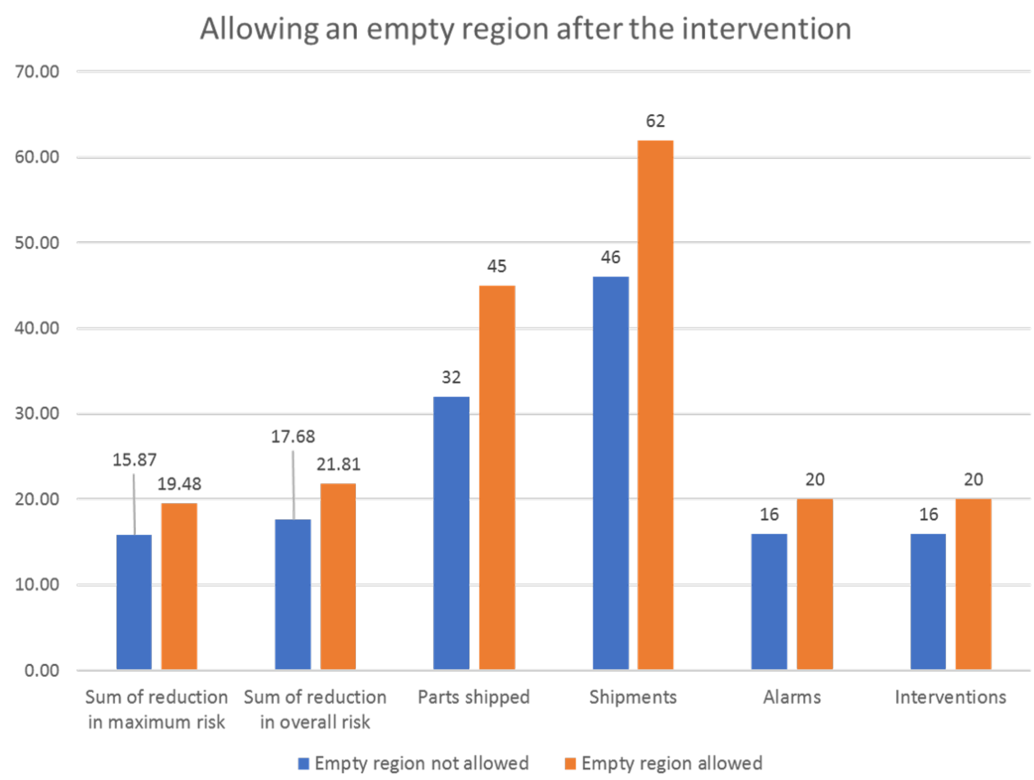

Figure 4.6: Comparison between allowing and disallowing an empty region after an intervention

Intervention handling interval

To analyze how often the alert and intervention generating model should be used, I generated alerts based on historic data from every Monday in 16 weeks in May, June, July and August. In figure 4.7 I compare a monthly interval with a weekly interval. The orange bar shows the number of 12NCs that have had an alert in a week that the tool is analyzed, but did not have an alert in the previous or next week that the tool is analyzed. The blue bar shows the number of alerts on the day that the alerts are generated. The number of missed 12NCs is relatively large and therefore I recommend to rerun the analysis at least weekly.

Figure 4.7: Number of (Missed) 12NCs with an alert (monthly interval vs. weekly interval)

If alerts are analyzed weekly, it is still possible that alerts are being missed. However the chances are smaller since the demand for a spare part is almost always below one part per month and supplier lead times are in the order of several months.

4.4 Conclusion

In this chapter I have introduced the two-staged alert and intervention generating model that proposed proactive lateral transshipments to minimize and balance the risk on a non-availability between the regions when the Central Warehouse is unable to replenish the shortages at the Local Warehouses.

The alert generating model generates an alert when there is at least one region eligible to receive support and one region eligible to give support. The expected unplanned non-availabilities of the receiving region need to be at least 0.3 and for the supporting region not larger than 0.5. The scheduled receipts to the warehouse can maximally replenish 40% of the shortages in the warehouses eligible to receive support.

The intervention generating model minimizes the maximum regional risk by proposing lateral transshipments. An alternative formulation minimizes the overall risk, however I have shown that the first formulation of the objective function is preferred since it is focused on achieving a balanced network. ASML does not want that the supporting region ships all its parts to other regions. I have shown that removing this constraint can give even better performance. Therefore I recommend the exception handling analyst to allow emptying the supporting region when the impact is substantial and the intervention is feasible. Lastly I recommend to refresh the alert generating model at least once a week to avoid missing alerts and intervention opportunities that could prevent non-availabilities.

Chapter 5

Results and impact

This section will explain the results that can be obtained by implementing the two-staged alert and intervention generating model for proactive lateral transshipments. The first section describes the numerical results that have been found by running the model on historic data. The second section describes the impact of the proactive lateral transshipments on the unplanned non-availabilities and international emergency shipments. The third section compares the costs saved with the costs of transshipping parts and finally an update of the implementation status is given.

5.1 Numerical results

Alerts and interventions per week

The alert and intervention generating models are tested by using data from every Monday in a 16 week period between May and August 2018. Figure 5.1 shows the number of alerts generated every week. In this figure is that in a few cases, an alert is generated for a situation in which an intervention is not preferred (the gray blocks). From this data can be found that when no interventions are executed, on average each week 20.6 12NCs face an unbalanced situation, as defined by the decision rules described in section 4.2.1.

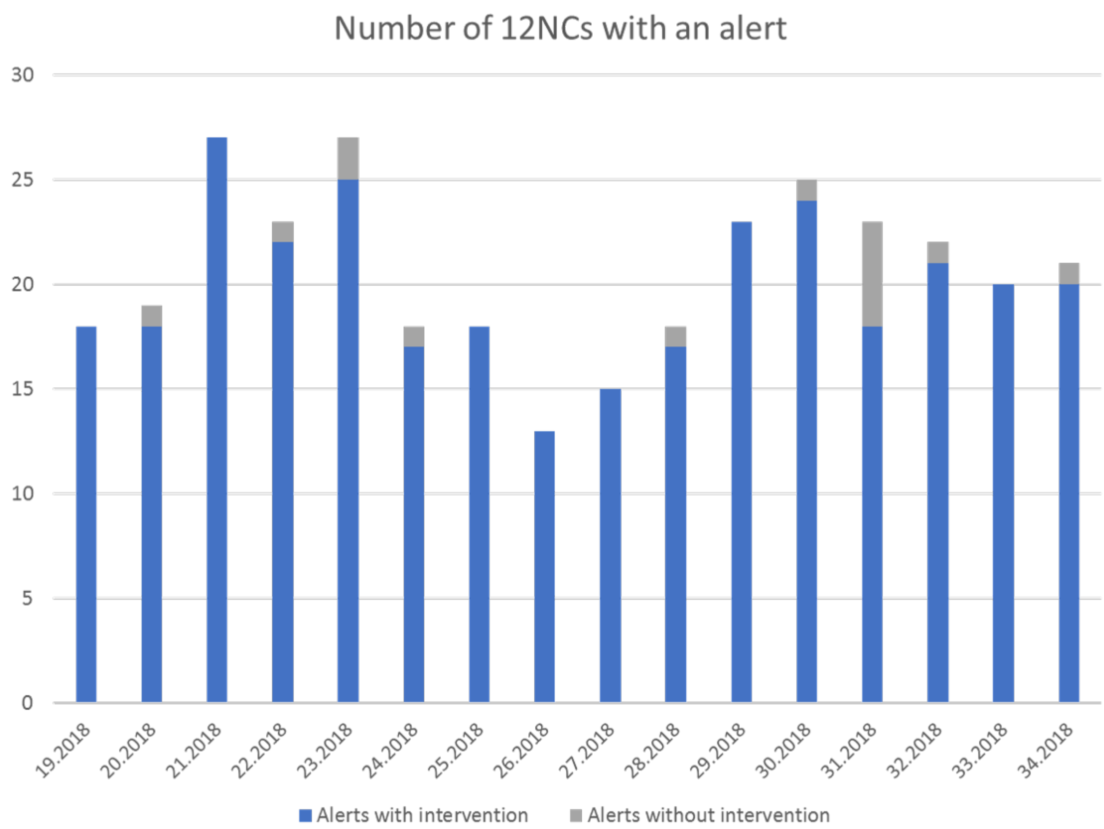

Figure 5.1: Number of 12NCs with an alert

The goal of this Control Tower intervention, is to balance the risks of the regions in the network. Therefore it is assumed that when an intervention is executed, an alert will not be generated for the specific 12NC. The effect on the number of alerts is shown in figure 5.2. After an initialization period of one month, the average number of alerts per week (read: number of 12NCs with an unbalanced network) will decrease to 8.5. This is a reduction of 58.7%

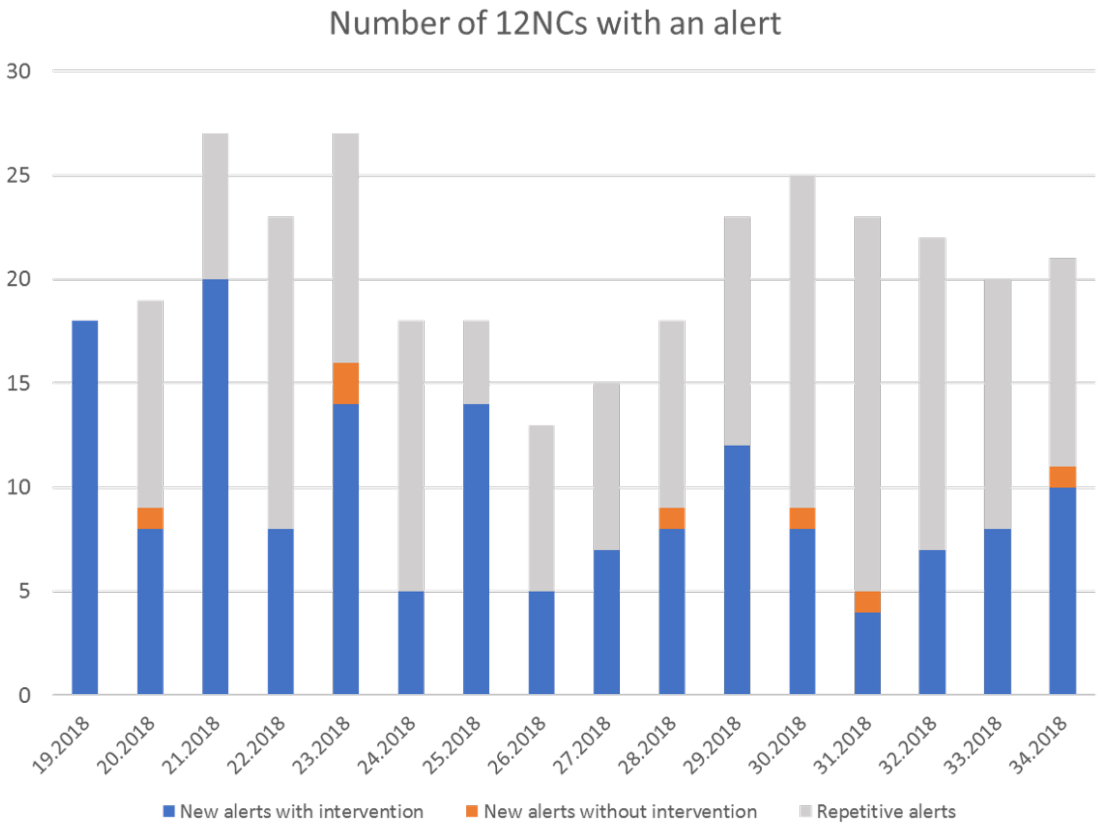

Figure 5.2: Number of new 12NCs with an alert

Impact on availability

The average impact on the maximum regional expected unplanned non-availabilities is calculated by using only new alerts after the initialization period of one month. The average maximum regional expected unplanned non-availabilities in the next month before the intervention proposal is 1.31. The intervention proposed by the model reduce the expected unplanned non-availabilities in the next month to 0.21, a reduction of 1.10 unplanned non-availabilities. This is an reduction of 84%.

Orderlines and parts shipped

An intervention proposal can consist of multiple transshipments and a transshipment can contain multiple parts. Figure 5.3 shows the number of orderlines and shipped parts in each week. In this chart only interventions for *new* alerts are taken into account. The average number of transshipped parts is 19.5 and the average number of orderlines is 13.8. This makes sense since most often, only one part is shipped from one region to another. The financial implications of lateral transshipments will be discussed later in section 5.3.

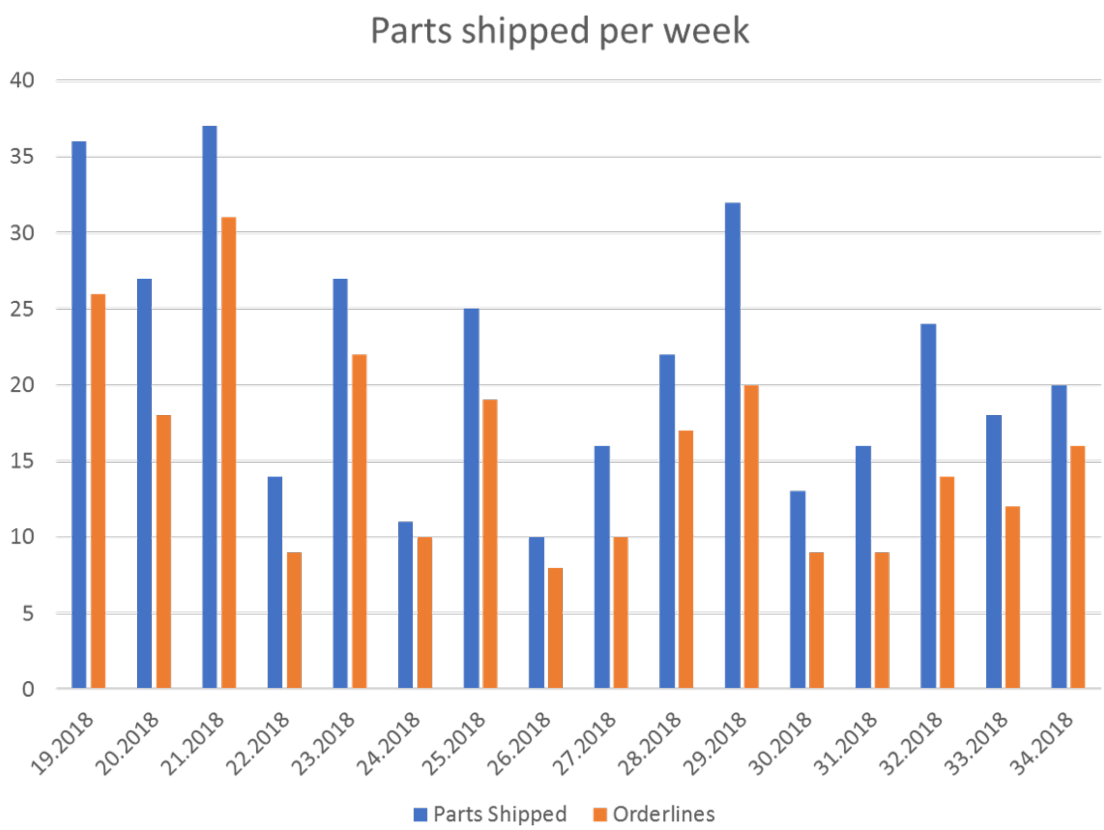

Figure 5.3: Number of orderlines and parts shipped for 12NCs with a new alert

Breakdown of alerts in usage classes

In the next cross-sections, I look at the types of the 12NCs that are being alerted. In figure 5.4 you can see the breakdown of the number of alerts per week into usage categories. Table 5.1 shows the thresholds for the three categories. In this graph we see that the largest portion of the alerts are for fast moving parts. This bucket is the smallest bucket in the total number of spare parts, however, when compared to the unplanned non-availabilities in the same 16 week period, the percentages are quite comparable. This makes sense since slow moving parts have more likely to be replenished in time and therefore naturally have lower risks. Using different parameter settings for slow movers with less strict thresholds for the risks or scheduled receipts, will lead to more alerts and interventions for slow moving parts, however the impact of these interventions will be small since the risks are lower and most likely will not increase the number of prevented unplanned non-availabilities.

Table 5.1: Usage categories

Usage category	**Bucket size**	**Bucket size**
	service parts	unplanned NAVs
Slow mover	70%	23%
Medium mover	16%	13%
Fast mover	14%	63%

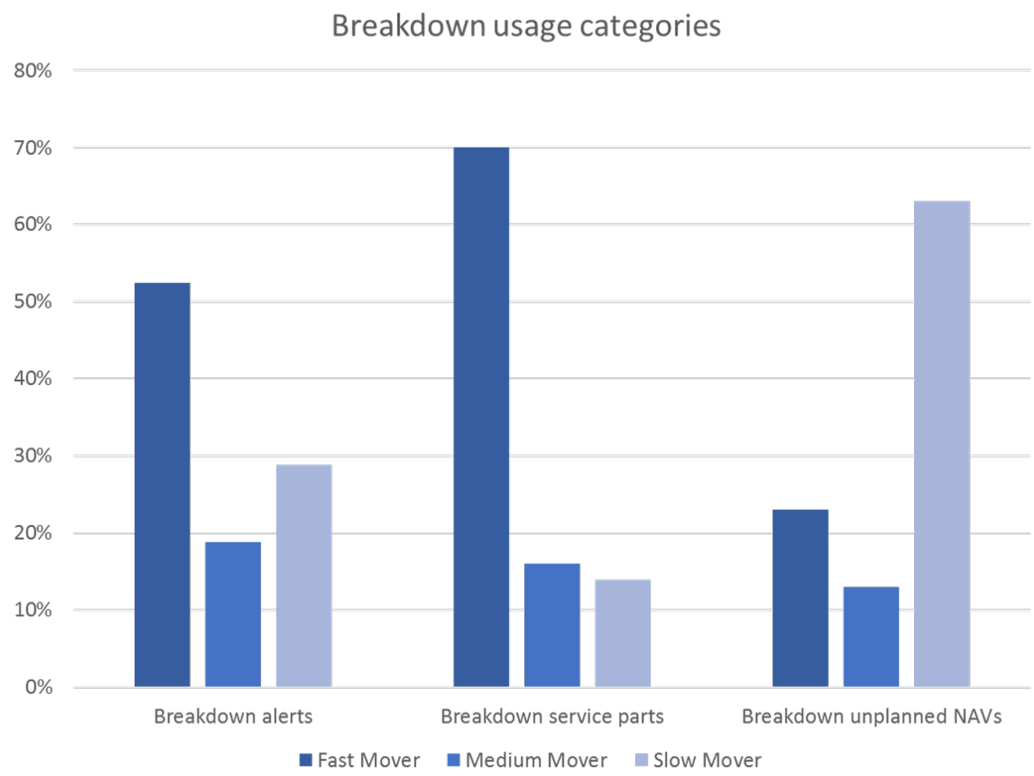

Figure 5.4: Number of 12NCs with an alert per usage category

Please note that NORA uses another formula to approximate the expected unplanned non-availabilities in the next month for parts at local warehouses and regions where they have no forecast. In those cases, NORA uses the BSL when applicable as forecast. A part will have no forecast on local warehouse and regional level, when the part has not had usage of more than ■ parts since it has been introduced. In those cases the BSL is based on the initial failure rate indicated by the design & engineering department. This logic makes sense for the prioritization of replenishment, since the local warehouses and regions with the highest base stock level will be replenished first. However, it does not make sense to transship parts based on these risk, since the initial failure rate and thus the risk on a non-availability is commonly much lower in reality. Planners from the FMA department indicate that they would not consider an intervention as an initial replenishment for a region (the region did not have stock for this part earlier). Therefore, alerts are not generated for parts that do not have a monthly forecast at the local warehouse level.

Breakdown of alerts in cost classes

The second breakdown shows the distribution of alerts in cost classes. Figure 5.5 shows the distribution of the cost classes per week. Table 5.2 shows the boundaries of the cost categories and size of the category with respect to the number of service parts and with respect to the number of unplanned non-availabilities. The distributions cost classes in the 12NCs with an alert match with the distributions of the total population and the distribution of the cost classes in the set of 12NCs with an unplanned non-availabilties in the same period. Therefore, I don't propose to use different parameters for the different cost categories.

Table 5.2: Cost categories

Cost category	**Bucket size** service parts	**Bucket size** unplanned NAVs
Cheap	40%	33%
Low cost	31%	33%
Medium cost	18%	21%
High cost	9%	12%
Expensive	2%	1%

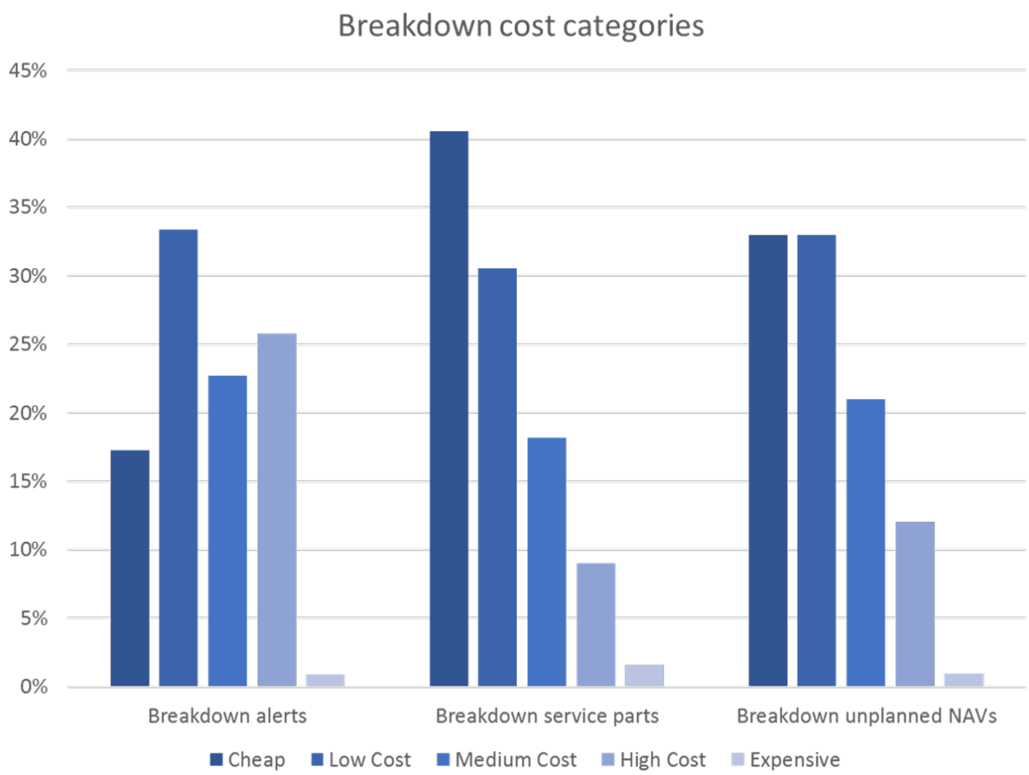

Figure 5.5: Number of 12NCs with an alert per cost category

Breakdown of alerts in machine types

The next breakdown shows the different machine types (also called: product platforms) to which the 12NCs that have an alert are dedicated to. The *PAS* systems are the oldest systems and most of the LPA contracts for machines on this platform are phased out. The DUV platform has the highest number of 12NCs that are dedicated to this platform and the highest installed base. This means that the usage of spare parts is the highest for this platform, hence the base stock levels are the highest for 12NCs applicable to this platform. EUV machines can use EUV specific parts, but also uses parts that can be found on DUV machines.

Figure 5.6 shows the number of alerts per machine type. It also shows the distribution of the machine types among the total group of spare parts and compares the unplanned non-availabilities per machine type. The EUV bars only show the parts specific to EUV machines. Usage for these parts is lower and it is common for regions to have a BSL of 1. Therefore, it is beneficial for the availability of these specific spare parts to neglect the restriction that the Inventory Position of the supporting region needs to be at least 2 and that a region cannot be empty after support.

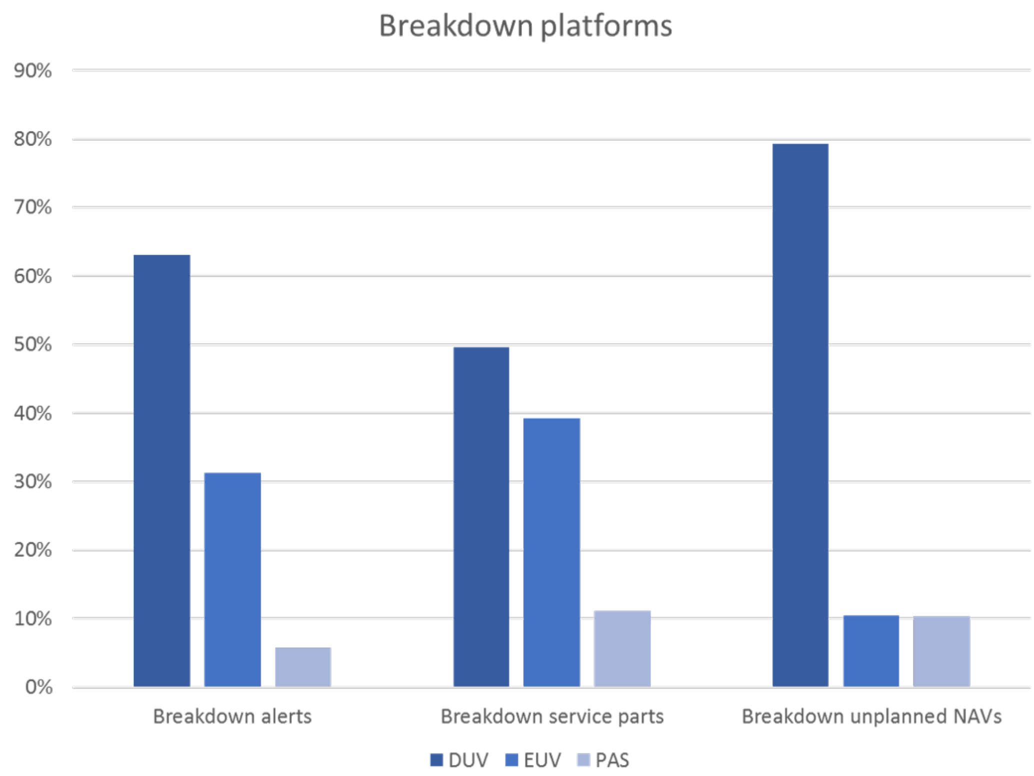

Figure 5.6: Number of 12NCs with an alert per machine type

Breakdown of alerts per region

The last breakdown shows the quantity of parts a region should have shipped and received in the 16 week period. For this analysis I only looked at unique 12NCs, assuming that the first intervention for the 12NC resolves the unplanned non-availability risk enough so that another intervention is not required. It becomes clear that the regions H and J receive the most support while region B and D need to send the most parts. This can be explained by the fact that region H and J are the largest regions with respect to usage and the height of the base stock levels.

Regions B, C and D are regions within the same country. Due to legislative reasons, it is harder and more expensive to ship parts between the warehouses in this country. This results in relatively higher base stock levels due to less risk sharing between the local warehouses in this country. Moving parts from this country to another country is less complicated and less expensive.

Using different parameters for the different regions is not preferred, since the goal is to balance the network and equalize the expected unplanned non-availabilties in the next month for each region.

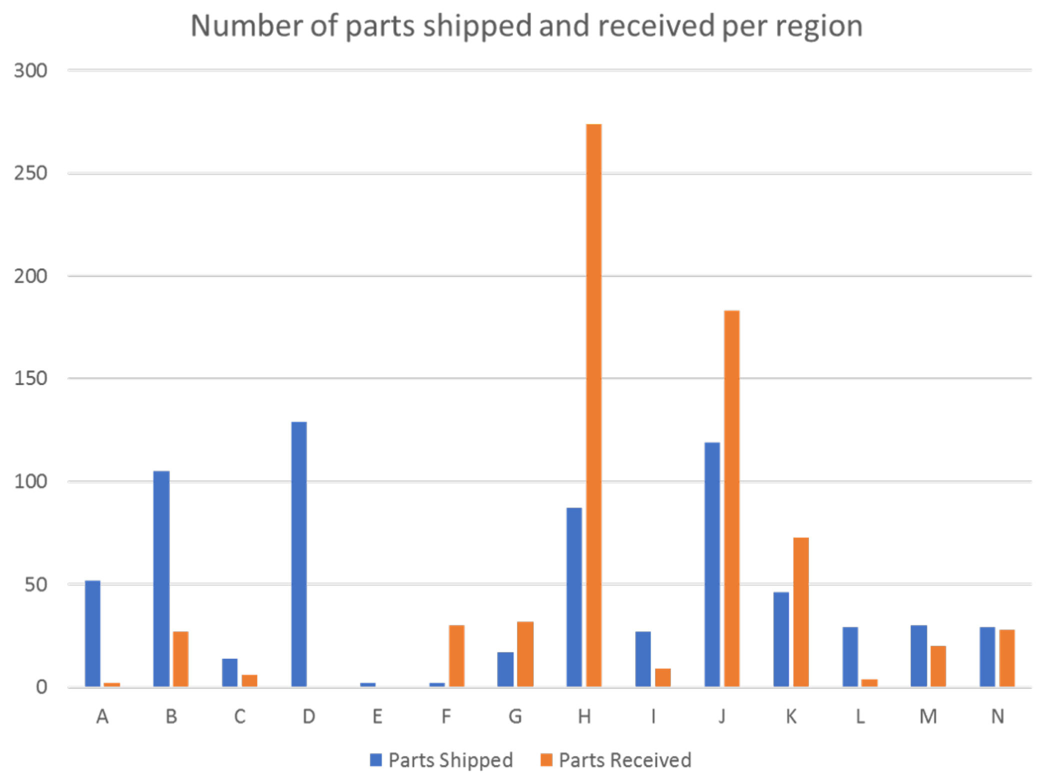

Figure 5.7: Number of parts sent and received per region for new alerts

Please note that in this analysis, the trends are analyzed by only looking at the first intervention of a 12NC. Using the alert and intervention generating model to adjust base stock levels would be possible if the number of interventions executed would be monitored over time. When the same kind of transshipment is proposed multiple times it could indicate that the base stock levels have to be adjusted.

5.2 Impact on unplanned non-availabilities

The Control Tower process at ASML seeks to find and resolve exceptions in the Supply Chain that could eventually lead to a non-availability. With the alert and intervention generating model I try to detect imbalanced inventories and try to resolve these with lateral transshipments between regions. To see how much unplanned non-availabilities could be avoided by this intervention, I have analyzed previous unplanned non-availabilities that occured between May and August 2018.

For this analysis I use the data set in which Emergency and Priority shipment are logged by the emergency handling department. Corresponding with the definition of unplanned non-availabilities in section 1.7.2 I filtered the data on: 1. Emergency shipments only, 2. Shipped part is used for a machine down, 3. The shipped part is used, and 4. The Emergency shipment is sent to a local warehouse that has a BSL for the part. Then I matched the set of unplanned NAVs with the set of alerts that were generated in the same period. The results can be shown in figure 5.8 and will be discussed after the chart.

Figure 5.8: Number of 12NCs with an unplanned NAV

This figure is removed in the version for publication.

Figure 5.8 shows the unplanned non-availabilities per week for which no alert was generated in four shades of grey. The shade of gray shows from which other warehouse the part has been shipped to serve the customer. Non-availabilities are supported from four sources: 1. the Central Warehouse, 2. the Emergency Hub, 3. another Local Warehouse in the same region, or 4. another Local Warehouse from another region. The largest portion of unplanned non-availabilities are supported from another local warehouse in the same region. The reason why the majority of these NAVs cannot be seen by the alert generation model is that the expected unplanned non-availabilities in the region is sufficient, however the parts are on stock on the wrong locations within the region. This phenomenon can be attributed to the SPartAn algorithm that allocates the base stock levels to local warehouses. The logic in this algorithm includes the possibility of emergency transshipments within the region. A risk pooling effect can be, achieved reducing the base stock level in the local warehouses. In those cases where an alert is generated for NAVs supported from within the region, the network is imbalanced and proactive Lateral transshipments can improve the availability in that region.

When the non-availability can not be solved by using an emergency from within the region, the next preferred solution is an emergency shipment from the central warehouse. When new supply arrives at the central warehouse, it is immediately used to replenish the local warehouses. Some cases are detected where an alert was generated for a NAV which was eventually supported from the central warehouse. In these situations, the alert was generated weeks in advance when there was no stock at the central warehouse and no scheduled receipts within two weeks. Preventing an emergency from the central warehouse by balancing the risks in the downstream network is preferred. The part can than be used for replenishing the region with the highest risk.

The next preferred transshipment is from either another region or from the emergency hub. The worldwide support team will determine which solution is used. Avoiding these emergencies is desirable since the emergency will take longer and the escalation costs and transshipment costs are the most expensive. Unfortunately, this can only be avoided when there is a region eligible for support or the emergency hub has a part on stock.

Table 5.3 shows the approximate number of unplanned non-availabilities that can be avoided on a yearly basis.

Table 5.3: Approximate number of unplanned non-availabilities avoided per year

This table is removed in the version for publication.

Since there are more emergency shipments than only the unplanned non-availabilties, I have also looked at the number of 12NCs with an emergency that had an alert in the period before. Please note that when a local warehouse does not have a BSL, the risk on an *unplanned* non-availability is zero and therefore the expected unplanned non-availabilities in the month for this local warehouse is also 0. Therefore very few other emergency shipments have had an alert before the shipment itself. Table 5.4 shows the approximate number of Emergency shipments that can be avoided on a yearly basis.

Table 5.4: Approximate number of emergencies avoided per year

This table is removed in the version for publication.

Please note that after an intervention, the risk in the supporting region increases. Therefore, there exists a chance that the transshipment proposed by the intervention generating model creates a non-availability in the supporting region, and not in the region to which it supported to. To arrive at a more accurate estimate on how much non-availabilities are avoided and causes by the proactive lateral transshipments, a simulation study would be beneficial. Due to time limitations, this is a recommendation for further research.

5.3 Economic impact

This section is removed in the version for publication.

5.4 Implementation

During the development of the model, the CSCM department staff expressed their interests in the corresponding Spotfire tool in which the decision rules are programmed. During the validation of the tool, the FMA department (part of CSCM) stated that they would like to use the tool as soon as possible. I am not directly involved in designing and implementing the processes and assigning responsibilities, however, the concerns the DALs (working within the FMA department) addressed during the validation of the tool were considered during the development. Before the first proactive transshipment can take place, the CSCM staff needs to give green light and the local warehouses need to be informed on this new way of working.

5.5 Conclusion

In this section I have showed that proactive lateral transshipments can be used in on average 20.6 cases per week. When the intervention is used, the number of alerts will drop to around 9.5 parts per week. The maximum regional expected unplanned non-availabilities will improve with 84%.

I have shown a breakdown of the number of alerts into different categories: 1. usage classes, 2. cost classes, 3. product platform, and 4. region. The results show that the same parameter settings can be used for all the parts and regions. A nuance can be made for the EUV platform, for which most of the parts are low usage (meaning low regional BSLs) or have unreliable forecast (meaning a too optimistic risk calculation).

On a yearly basis, approximately 38 unplanned non-availabilities will be avoided by proactive lateral transshipments. Approximately 78 international emergency shipments will be avoided by the proactive lateral transshipments. The saved costs of emergency shipments will save costs, but saved transportation costs alone are not enough to break even with the added transshipment costs. The penalty costs of not meeting a service contracts are not taken into account with this analysis.

Chapter 6

Conclusions and recommendations

In the first section of this chapter, I will give answer to the research questions I have stated in chapter 1. Next, the limitations and assumptions are discussed in the eponymous section. Lastly, I share recommendations for further research on the topic of proactive Control Towers and give directions for implementation and improvement to ASML in the last section.

6.1 Conclusions

The main research goal is to answer the following question: *How can prospective unplanned non-availability issues of after-sales service parts be recognized, analyzed and pro-actively acted upon, to increase the spare part availability of these parts on an operational level?*

The short answer to this question is: *By generating alerts when the availability is imbalanced across regions and proposing proactive Lateral transshipments as interventions.* The next subsections answer the subquestions in more detail.

6.1.1 Identifying non-availabilities in advance

In chapter 2 I describe an extensive analysis into the root causes of non-availabilities. These can be classified in five different categories: 1. Demand related causes, 2. Quality related causes, 3. Supplier related causes, 4. Stock related causes and finally 5. Tactical parameter related causes. The existing Control Tower triggers does already have detailed SCV capabilities with regards to the demand and supplier related causes. With regards to quality and tactical level related causes, the impact of added visibility would be very small compared to the required effort. Therefore, I've chosen to focus on improving the visibility of Stock related causes for a non-availability. To achieve this added visibility, I have developed decision rules to trigger alerts on a regional level, instead of the global level in the original Control Tower.

6.1.2 Impact on literature

This research has contributed to the existing literature of Control Towers by introducing a new two-staged decision model for proactive lateral transshipments. These lateral transshipments are triggered by the expected unplanned non-availabilities that express the risk on failing to serve the customer demand from the dedicated local warehouse. The alert generating model will give the Control Tower analyst a signal when a situation of imbalance is detected for a specific stock keeping unit. After that, an intervention is generated by the second stage model to transship parts to minimize risk in the region with the highest risk.

My contribution to the ProSeLoNext project is an example on how to manage automated processing of exception messages (alerts) and using advance supply and demand information to plan operational interventions proactively. These two points were brought up by the members of the consortium as promising research directions.

6.1.3 Decision rules for alert generation

The Proactive Regional Support Tool consists of two stages: 1. the alert generation model in the *Data Application Layer* of the Control Tower, and 2. the intervention generation model located in the *Operational Planning Layer* of the Control Tower. An alert is received when there is at least one region with relatively high risk on a non-availability, and at least one region with relatively low risk on a non-availability, such that a proactive Lateral transshipment can be executed. The requesting region must have at least 0.3 expected unplanned non-availabilities per month while the supporting region can not have more than 0.3 expected unplanned non-availabilities per month. Another restriction is that the central warehouse must be unable to replenish more than 40% of the shortages in the downstream local warehouses within two weeks. Lastly, the supporting region should always have one part remaining on stock after it supported. Therefore it should have at least two parts as on-hand inventory.

6.1.4 Generating interventions

An intervention generating model is used to propose interventions for the 12NCs with an alert. These interventions can be interpreted and approved by the Control Tower analyst. The objective function of the model is to minimize the maximum regional expected unplanned non-availabilities per month for every individual alert. This means that when a part is transshipped from one region to another, the expected unplanned non-availabilities in the receiving region will decrease while it will increase in the supporting region. When the receiving region was the region with the highest expected unplanned non-availabilities per month, the objective function will decrease. The regional expected unplanned non-availabilities in the next month are calculated in advance for a predefined range of Regional inventory position changes to cope with the non-linear function of the expected unplanned non-availabilities. The decision variables of this linear programming model are binary variables for every region and Inventory Position change to indicate if and how many parts a region has to send or receive.

Constraints to the model are: 1. The Regional inventory position after the intervention is the initial Inventory Position plus or minus the proposed change; 2. The quantity of the parts sent must be equal to the quantity of parts received; 3. A decision on the Inventory Position of a region after the interventions must be made; 4. The Inventory Position can only decrease in a region when the region is eligible to *give* support; 5. The Inventory Position can only increase in a region when it is eligible to *receive* support; 6. At least one part remains in the supporting region.

Another option for the objective function is to minimize the sum of regional expected unplanned non-availabilities in the next month. This formulation is not chosen, because in essence it does not solve the imbalanced network. In practice, the performance of the two models do not deviate largely from each other. The formulation that minimizes the maximum risks has slightly less maximum regional expected unplanned non-availabilities, while the objective function that minimizes the sum has a slightly lower overall risk.

One of the constraints that could make the intervention proposal sub-optimal from a risk minimization perspective, is the decision to disallow emptying the supporting regions to an Inventory Position of zero parts. This is decision is made by the staff and DAL planners that will use the model and decide to execute the interventions. Removing the constraint that a region needs to have at least two parts to be eligible for support yields 4 more alerts in the test set (increase of 25%). Removing the constraint in the intervention generating model, results in an increase of the number of shipments of 40% and a decrease in the maximum regional unplanned non-availabiltilies in the next month of 22.7%.

6.1.5 Impact of the two-staged alert and intervention generating model

With the Proactive Regional Support Tool, an alert will be generated in approximately 20 situations per week. This will drop to 9.6 new alerts per week when an intervention is executed and results in a better balanced network. The impact of this intervention on the maximum regional expected non-availabilities per month is a reduction of 84% on average. Furthermore, it is expected that around 38 unplanned non-availabilities can be avoided by proactively transshipping parts.

The tool should be used once a week to generate alerts in order to achieve a high level of SCV with acceptable requirements on computation power and manual analysis. The dashboard contains all the information a material availability planner would need for decision making including suggestions for interventions.

6.2 Discussion

There are a few limitations I have faced during this research, namely, 1. the calculation for expected unplanned non-availabilities is not the best approximation for this risk and 2. the intervention generation model does not include contract performance and transshipment costs in the model. The following subsections will elaborate on these points.

6.2.1 Calculation of expected unplanned non-availabilities

In this research, I have used a calculation of expected unplanned non-availabilities in the next month as primary variable to generate alerts and to optimize in the intervention generating model. A major issue with this calculation is the way the expected value is calculated. The procedure multiplies the Blocking Probability with the Forecast for the usage in the next month which is statistically incorrect. A better way to calculate the expected value is the calculation of Expected Backorders, which is given as well. When comparing these two calculations, I discovered that there is a strong positive correlation between the two calculations. This means that when there are many expected unplanned non-availabilities, there are also many expected backorders and vice versa. The main difference between the two calculations is the fact that the expected unplanned non-availabilities are roughly two times larger than the expected backorders. This means that the calculation for the expected unplanned non-availabilities can be used to indicate relative risks (a high value means high risk and a low value means a low risk, but the numbers do not represent the real expectation. See appendix B for more context.

Nonetheless, I chose to use the expected unplanned non-availabilities in the next month because there are some differences between the two calculations in the resulting priorities. Where the expected backorder calculations gives an almost negligible difference between two values, the expected unplanned non-availabilities calculation gives a much larger range of priorities. In those cases it is important that the Proactive Regional Support Tool shares the same business logic and priorities with the other activities at ASML (e.g. automated replenishment and the development of a new Service Level Agreement (SLA)).

6.2.2 Economical optimization opportunities

The focus of the Control Tower is maximizing the availability of spare parts for the customer. However, the economical impact of an intervention could be considered when proposing interventions. For instance, the costs of a transshipment partially depends on the distance between the sending and receiving region. Therefore, a shipment from the closest region with slightly more expected unplanned non-availabilities could be preferred over a shipment from a distant region that has very little expected unplanned non-availabilities. At the moment, analyzing this trade-off is undesired by ASML.

Furthermore, the contract performance is not considered with this Proactive Regional Support Tool. A proactive lateral shipment from a region where the risk of breaching the contracted availability is high is not preferred. This is because the large penalty costs ASML has to pay when the SLA can not be met.

6.3 Recommendations

Finally, I have suggestions for further research and recommendations for the implementation of proactive decision making in the Control Tower of ASML.

6.3.1 Further research

The impact on prevented Emergency shipments and unplanned non-availabilities can only be analyzed afterwards on historic data. These numbers contain uncertainty, because when an intervention is executed, the network has changed and other decisions might be made. I performed case studies for a few 12NCs to see the impact of an intervention. This requires gathering and analyzing accurate and detailed data on usage and inventory positions in the network over a long time period. Using proactive Lateral transshipments in practice can not give conclusive numbers on prevented non-availabilities as well, since it is unclear what decisions were made if the model was not in use.

A reliable method to test the impact of the decision rules and interventions would be simulating the network with and without proactive Lateral transshipments. A simulation study with a validated supply chain network with similar characteristics can give more reliable performance indicators and enables testing the impact of decisions like transshipment times (Routine or Priority) and whether or not to allow empty regions after supporting others.

6.3.2 Implementation

The highest level of SCV can be obtained when the Proactive Regional Support Tool is used at least once a week. For this, the same input data from NORA can be used, in combination with an output file containing the expected unplanned non-availabilities. The intervention proposals are input as well. These have to be generated first by running the accompanying Excel tool. I suggest three topics for further improvement: 1. logging the executed intervention with a dedicated administrative key, such that statistics on performance can be gathered, 2. improving the robustness of the tool and 3. implementing cost parameters and contract performance in the decision rules.

Enable continuous improvement and automation I recommend to developing an automated process that can monitor the network of a 12NC for which an intervention has been executed. This enables reliable determining the impact of categorized interventions on the long term. It can also be used for regular analysis of the performance of the parameter settings.

Furthermore, monitoring the impact of interventions enables the possibilities for further automation using machine learning, where the network is analyzed continuously and interventions are executed based on the expected impact that is accurately determined by previous interventions.

Robustness of the Proactive Regional Support Tool With robustness I mean the reliability and functionality of the tool. Data preparation and loading into the dashboard could take less time when the data is pulled directly out of SAP (Information Perception Layer) and the calculations for expected unplanned non-availabilities per month per region are done within the Control Tower architecture (in the Operational Data Storage layer). This will make the Proactive Regional

Support Tool independent of the NORA tool. Furthermore, the functionality of the Excel tool can easily be damaged by (unknowingly and unwillingly) changing the wrong cell. Building a proper decision support application that is easy and fast to use would be a more robust solution.

Cost parameters and contract performance Currently, the Control Tower analyst has to approve all the proposed interventions. All relevant information of the situation in the Supply Chain is visible in the tool, however the decision is also dependent on cost parameters and contract performance. When the contract in a region is expected to be breached, supporting parts from that region would not be executed. Furthermore in economical perspective, it can be better to transship two parts from a region than to transship one part from two regions. The intervention generating model does not have a preference when two regions have equal risks after supporting, while the nearest region is definitely better in economical sense. ASML should prioritize sharing reliable information across departments, such that better decision making is enabled.

Other operational interventions The study of Topan and Van der Heijden (2018) show that the Proactive Lateral Transshipments can effectively be combined with Proactive Emergency Shipments from an Upstream Warehouse. An implementation project on the latter operational intervention could yield even more availability improvements.

References

Alstrøm, P., & Madsen, P. (1996). Tracking signals in inventory control systems a simulation study. *International journal of production economics*, *45*(1-3), 293–302.

ASML. (2017). *Wafer treatment process.* Website ASML Press image library. Retrieved from https://www.asml.com/press/image-library/en/s44169

ASML. (2018a, February). *What we do - asml.* ASML Website. Retrieved from https://www.asml.com/company/what-we-do/en/s277?rid=51981

ASML. (2018b, February). *Why we exist - asml.* ASML Website. Retrieved from https://www.asml.com/company/why-we-exist/en/s277?rid=51983

Axsäter, S. (2003). A new decision rule for lateral transshipments in inventory systems. *Management Science*, *49*(9), 1168–1179.

Banerjee, A., Burton, J., & Banerjee, S. (2003). A simulation study of lateral shipments in single supplier, multiple buyers supply chain networks. *International Journal of Production Economics*, *81*, 103–114.

Basten, R. (2016, November). *News university of technology eindhoven.* Website University of Technology Eindhoven. Retrieved from https://opac.ieis.tue.nl/proactive-service-logistics-next-has-been-awarded-a-one-milion-euro-grant/

Bleda, J., Martin, R., Narsana, T., & Jones, D. (2014). *Prepare for takeoff with a supply chain control tower.* Accenture.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, *184*(3), 1140–1154.

Chew, E. P., Lee, L. H., & Liu, S. (2013). Dynamic rationing and ordering policies for multiple demand classes. *OR spectrum*, *35*(1), 127–151.

Enders, P., Adan, I., Scheller-Wolf, A., & van Houtum, G.-J. (2014). Inventory rationing for a system with heterogeneous customer classes. *Flexible services and manufacturing journal*, *26*(3), 344–386.

Evans, R. V. (1968). Sales and restocking policies in a single item inventory system. *Management Science*, *14*(7), 463–472.

Evers, P. T. (2001). Heuristics for assessing emergency transshipments. *European Journal of Operational Research*, *129*(2), 311–316.

Fadıloğlu, M. M., & Bulut, Ö. (2010). A dynamic rationing policy for continuous-review inventory systems. *European Journal of Operational Research*, *202*(3), 675–685.

Foster Jr, S. T. (2008). Towards an understanding of supply chain quality management. *Journal of operations management*, *26*(4), 461–467.

Gaukler, G. M., Özer, Ö., & Hausman, W. H. (2008). Order progress information: Improved dynamic emergency ordering policies. *Production and Operations Management*, *17*(6), 599–613.

Gesing, B., Peterson, S. J., & Michelsen, D. (2018). *Artificial intelligence in logistics* (Tech. Rep.). IBM and DHL.

Glazebrook, K., Paterson, C., Rauscher, S., & Archibald, T. (2015). Benefits of hybrid lateral transshipments in multi-item inventory systems under periodic replenishment. *Production and Operations Management*, *24*(2), 311–324.

Govindan, K., Cheng, T., Mishra, N., & Shukla, N. (2018). *Big data analytics and application for logistics and supply chain management.* Elsevier.

Grahovac, J., & Chakravarty, A. (2001). Sharing and lateral transshipment of inventory in a supply chain with expensive low-demand items. *Management Science*, *47*(4), 579–594.

Hemeimat, R., Al-Qatawneh, L., Arafeh, M., & Masoud, S. (2016). Forecasting spare parts demand using statistical analysis. *American Journal of Operations Research*, *6*(02), 113.

Hermann, M., Pentek, T., & Otto, B. (2016). Design principles for industrie 4.0 scenarios. In *System sciences (hicss), 2016 49th hawaii international conference on* (pp. 3928–3937).

Hoadley, B., & Heyman, D. P. (1977). A two-echelon inventory model with purchases, dispositions, shipments, returns and transshipments. *Naval Research Logistics Quarterly*, *24*(1), 1–19.

Hoda Sabeti, O. A.-S., & Jaridi, M. (2016, July). Forecasting system monitoring under non-normal input noise distributions. *Journal of Industrial Engineering and Management*, *5*(2).

Howard, C., Marklund, J., Tan, T., & Reijnen, I. (2015). Inventory control in a spare parts distribution system with emergency stocks and pipeline information. *Manufacturing & Service Operations Management*, *17*(2), 142–156.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679–688.

Janakiraman, G., Seshadri, S., & Sheopuri, A. (2014). Analysis of tailored base-surge policies in dual sourcing inventory systems. *Management Science*, *61*(7), 1547–1561.

Jiang, C., & Sheng, Z. (2009). Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system. *Expert Systems with Applications*, *36*(3), 6520–6526.

Johansen, S. G. (2018). Emergency orders in the periodic-review inventory system with fixed ordering costs and stochastic lead times for normal orders. *International Journal of Production Economics*.

Johansen, S. G., & Thorstenson, A. (2014). Emergency orders in the periodic-review inventory system with fixed ordering costs and compound poisson demand. *International Journal of Production Economics*, *157*, 147–157.

Kaplan, A. (1969). Stock rationing. *Management Science*, *15*(5), 260–267.

Kranenburg, A. (2006). *Spare parts inventory control under system availability constraints* (Doctoral dissertation, Department of Industrial Engineering & Innovation Sciences). doi: 10.6100/IR616052

Kukreja, A., & Schmidt, C. P. (2005). A model for lumpy demand parts in a multi-location inventory system with transshipments. *Computers & operations research*, *32*(8), 2059–2075.

Lee, Y. H., Jung, J. W., & Jeon, Y. S. (2007). An effective lateral transshipment policy to improve service level in the supply chain. *International journal of production economics*, *106*(1), 115–126.

Luzzini, D., Caniato, F., & Spina, G. (2014). Designing vendor evaluation systems: An empirical analysis. *Journal of Purchasing and Supply Management*, *20*(2), 113–129.

Maestrini, V., Luzzini, D., Maccarrone, P., & Caniato, F. (2017). Supply chain performance measurement systems: A systematic review and research agenda. *International Journal of Production Economics*, *183*, 299–315.

Maestrini, V., Maccarrone, P., Caniato, F., & Luzzini, D. (2018). Supplier performance measurement systems: Communication and reaction modes. *Industrial Marketing Management*.

McCrea, B. (2005). Ems completes the visibility picture. *Logistics Management*, *44*(6), 57–61.

Minner, S., Silver, E. A., & Robb, D. J. (2003). An improved heuristic for deciding on emergency transshipments. *European Journal of Operational Research*, *148*(2), 384–400.

OLAP Solutions. (n.d.). *Business intelligence.* Website OLAP Solutions. Retrieved from http://olap.com/learn-bi-olap/olap-bi-definitions/business-intelligence/

Olde Keizer, M., Teunter, R., & Veldman, J. (2017). Joint condition-based maintenance and invenotry optimization for systems with mulitple components. *European Journal of Operational Research*, *257*(1), 209-222.

Paterson, C., Kiesmüller, G., Teunter, R., & Glazebrook, K. (2011). Inventory models with lateral transshipments: A review. *European Journal of Operational Research*, *210*(2), 125–136.

Pinçe, Ç., Frenk, J., & Dekker, R. (2015). The role of contract expirations in service parts management. *Production and Operations Management*, *24*(10), 1580–1597.

Regattieri, A., Gamberi, M., Gamberini, R., & Manzini, R. (2005). Managing lumpy demand for aircraft spare parts. *Journal of Air Transport Management*, *11*(6), 426–431.

Rouse, M. (2009, October). *The essential guide to supply cahin management best practices.* Website TechTarget. Retrieved from https://searcherp.techtarget.com/essentialguide/The-essential-guide-to-supply-chain-management-best-practices

Sahin, M., Kizilaslan, R., & Demirel, Ö. F. (2013). Forecasting aviation spare parts demand using croston based methods and artificial neural networks. *Journal of Economic and Social Research*, *15*(2), 1.

Samuel, A. L. (1959, July). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. doi: 10.1147/rd.33.0210

Schwagmeier, R. (2016). *Operational control tower: For the after-sales service supply chain.* (Unpublished master's thesis). University of Twente.

Shou-Wen, J., Ying, T., & Yang-Hua, G. (2013). Study on supply chain information control tower system. *Information Technology Journal*, *12*(24), 8488.

Teunter, R. H., & Haneveld, W. K. K. (2008). Dynamic inventory rationing strategies for inventory systems with two demand classes, poisson demand and backordering. *European journal of operational research*, *190*(1), 156–178.

Tiacci, L., & Saetta, S. (2011). A heuristic for balancing the inventory level of different locations through lateral shipments. *International Journal of Production Economics*, *131*(1), 87–95.

Topan, E., Eruguz, A., Ma, W., van der Heijden, M., & Dekker, R. (2018). A review of operational spare parts planning for service control towers. *European Journal of Operations Research*.

Topan, E., & Van der Heijden, M. (2018). Operational level planning of a multi-item two-echelon spare parts inventory system. *European Journal of Operational Research*.

Topkis, D. M. (1968). Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science*, *15*(3), 160–176.

Trappey, A. J., Trappey, C. V., Fan, C.-Y., Hsu, A. P., Li, X.-K., & Lee, I. J. (2017). Iot patent roadmap for smart logistic service provision in the context of industry 4.0. *Journal of the Chinese Institute of Engineers*, *40*(7), 593–602.

Van Horenbeek, A., & Pintelon, L. (2015). A joint predictive maintenance and inventory policy. In *Engineering asset management-systems, professional practices and certification* (pp. 387–399). Springer.

Van Wijk, A., Adan, I. J., & van Houtum, G.-J. (2013). Optimal allocation policy for a multi-location inventory system with a quick response warehouse. *Operations Research Letters*, *41*(3), 305–310.

Vitasak, K. (2005). *Supply chain and logistics terms and glossary, council of supply chain management professionals, lombard, il.*

Wang, L., Chu, J., & Mao, W. (2008). An optimum condition-based replacement and spare provisioning policy based on markov chains. *Journal of Quality in Maintenance Engineering*, *14*(4), 387–401.

Wang, L., Chu, J., & Mao, W. (2009). A condition-based replacement and spare provisioning policy for deteriorating systems with uncertain deterioration to failure. *European Journal of Operational Research*, *194*(1), 184–205.

Washington State Department Of Enterprise Services. (n.d.). *Root cause analysis.* Website Washington State Department of Enterprise Services. Retrieved from https://des.wa.gov/services/risk-management/about-risk-management/enterprise-risk-management/root-cause-analysis

Wong, H., van Houtum, G.-J., Cattrysse, D., & Van Oudheusden, D. (2006). Multi-item spare parts systems with lateral transshipments and waiting time constraints. *European Journal of Operational Research*, *171*(3), 1071–1093.

Xie, J., & Wang, H. (2008). Joint optimization of condition-based preventive maintenance and spare ordering policy. In *Wireless communications, networking and mobile computing, 2008. wicom'08. 4th international conference on* (pp. 1–5).

Xu, L. D. (2011). Information architecture for supply chain quality management. *International Journal of Production Research*, *49*(1), 183–198.

Yang, G., Dekker, R., Gabor, A. F., & Axsäter, S. (2013). Service parts inventory control with lateral transshipment and pipeline stockflexibility. *International Journal of Production Economics*, *142*(2), 278–289.

Appendices

Appendix A

Calculation of criticality ranking

The criticality level is an awarded ranking assigned to all the spare parts in the ASML supply chain network. The most critical parts are parts that globally have little on-hand inventory in comparison to the base stock levels. Reasons for criticallity are explained in section 2.2.1.

The criticality levels are in order of importance:

- L3
- L2
- L1
- A
- B
- C
- D
- not critical

The ranking is determined by scoring all 12NCs on different categories. The weighted sum of these scores determine in to which criticality level a part belongs. The scoring categories are in order of importance:

1. Worldwide Fill-rate, defined as the ratio between on-hand inventory and the sum of all Base Stock Level (BSL) at the local warehouses and central warehouse
2. Expected Fill-rate in one month, by adding the scheduled receipts to the on-hand inventory and subtracting the forecasted usage of the part.
3. Scheduled upgrades and installs activities, since these activities can cause the need for a spare part
4. The history of non-availabilities for the part, scored for each region.
5. Manual input
6. Platform (recently introduced machines receive more weight than older platforms)

The process of scoring each 12NC to these criteria is repeated every month. This criticality level tirggers the Re-inning of new-buy and repair orders. It furthermore determines the priority in which the backlog of scheduling new-buy and repair orders is cleared, the backlog in which return shipments to higher echelon levels is cleared, the sequence in which materials with a Control Tower alert are analyzed.

Appendix B

Gap between expected non-availabilities and expected backorders

NORA calculates the $\mathbb{E}[NAV]$ daily, for each warehouse and item, to determine the replenishment priorities. The outcome of the alternative calculation shown in equation 4.6 is compared to the outcome of NORA. Figure B.1 shows the relation between the $\mathbb{E}[NAV]$ on the x-axis and the $\mathbb{E}BO$ on the y-axis. A straight line would mean a linear relationship between the two variables. This implies the NORA calculation can be used in practice since the prioritization between the two calculations is similar.

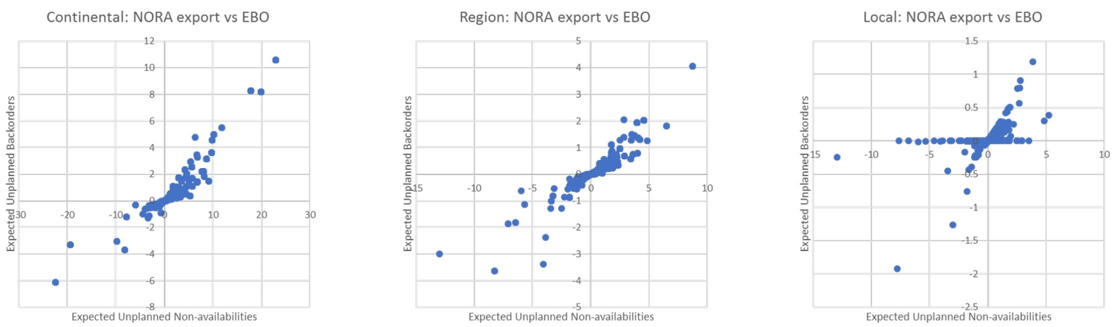

Figure B.1: The exported Expected Unplanned Non-availabilities versus the calculated Expected Backorders

In figure B.1 the comparison is made on continental, regional and local level. From the axis in the two graphs we can tell that the outcome of the $\mathbb{E}[NAV]$ calculation is too pessimistic in comparison to $\mathbb{E}BO$. For prioritization purposes, this would not necessarily be a bad thing, when the relationship between the variables is linear.

On the continental level, the relationship is almost linear. On the regioanl and local level, the two calculations deviate more. Part of the deviation is noise caused by locals with no forecast or base stock level. In these cases, the safety stock level is used as forecast, while the method in equation 4.6 returns just 0. Another example of different decision logic are exceptions for an emergency hub, where the base stock levels and inventory are adjusted by the algorithm.

When these exceptions are ignored, by calculating the $\mathbb{E}[NAV]$ manually, the resulting relationship for the local non-availability risk looks like figure B.2. In this figure a cubic curve is visible which includes 99.9% of the data points. 0.1% of the data points is found on a straight line intersecting the cubic curve in (0, 0). It is not clear why these two patterns occur, since the two curves share characteristics when compared to each other.

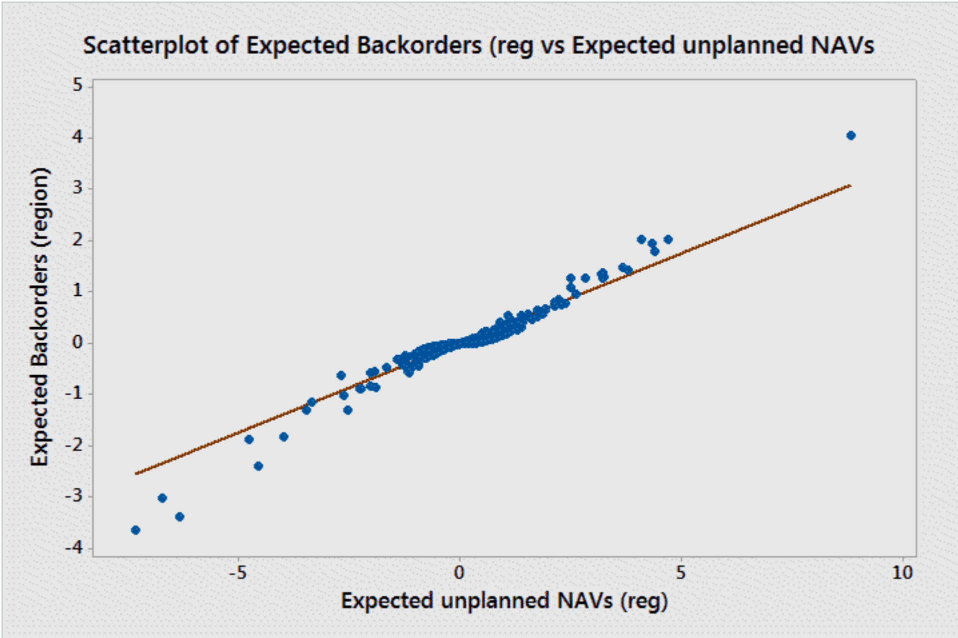

Figure B.2: Adjusted calculation of Expected Unplanned Non-availabilities versus the calculated Expected Backorders

An analysis of the correlation between the two calculations can be found in figure B.3. With a confidence level of 95%, the two calculations are significantly correlated. The Pearson coefficient of 0.955 indicates there is a strong positive correlation. This means when $\mathbb{E}[NAV]$ is high, $\mathbb{E}BO$ is high as well.

Correlation: Expected Backorders (region); Expected unplanned NAVs (reg)

```
Pearson correlation of Expected Backorders (region) and Expected unplanned NAVs (reg) = 0.955
P-Value = 0.000
```

Figure B.3: Result of correlation analysis

This significant correlation indicates that $\mathbb{E}[NAV]$ can be used to prioritize and assess the risk on a non-availability at a local warehouse or region, relative to other parts and regions.

Appendix C

Service orders

This appendix is removed in the version for publication.