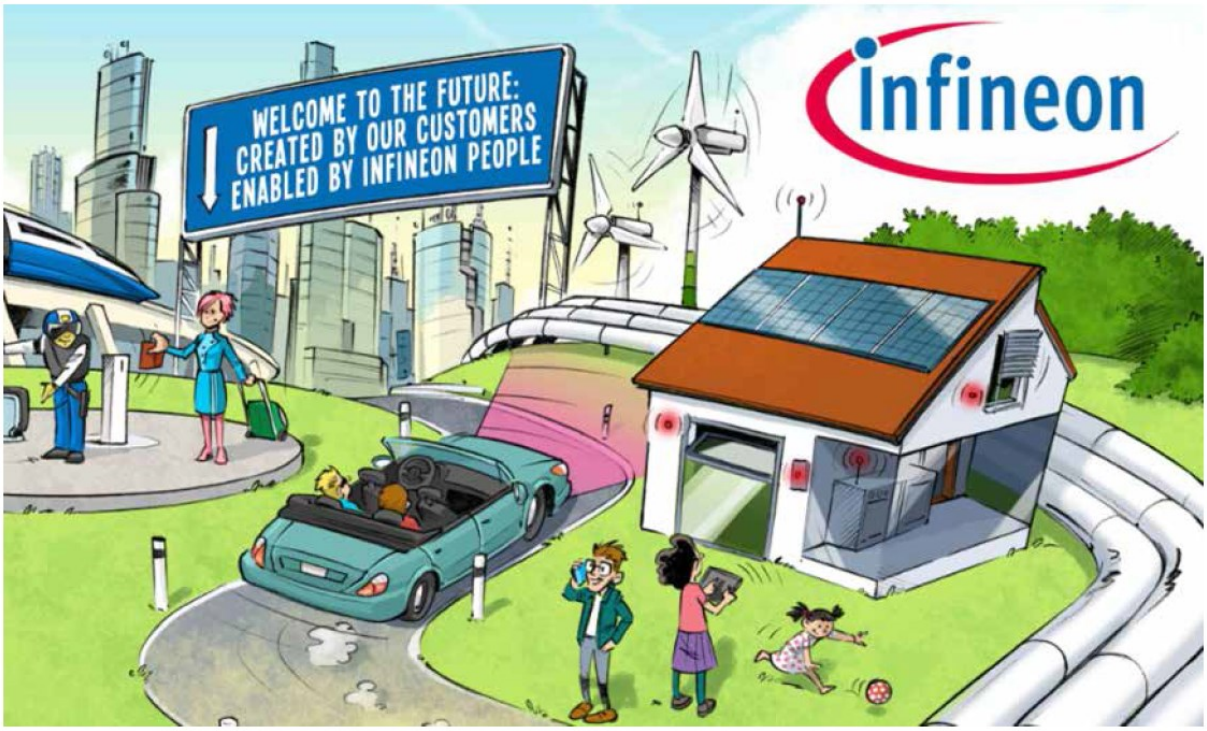
Balancing inventory and equipment contingencies

by a flexible semiconductor supply chain model



15.09.2015

Annemiek Adriaansen

UNIVERSITY OF TWENTE.

Project initiator	Infineon Technologies Am Campeon 1-12 85579 Neubiberg Germany www.infineon.com
Project title	Balancing inventory and equipment contingencies by a flexible semiconductor supply chain model.
Author	Anna M.A. (Annemiek) Adriaansen Master Industrial Engineering and Management Specialisation: Production and Logistics Management

Supervisory committee

Supervisors University of Twente	Dr. Ir. Ahmad Al Hanbali Dr. Ir. Marco Schutten
Supervisor Infineon Technologies	Sebastian Eirich, MSc.

Preface

After six months of hard work, I am proud to present you my master thesis. With this thesis I finalise my master in Industrial Engineering and Management and with that, my student days are over. Working-life is next and I am ready to discover what this new part in life will bring, in which I will be able to apply all the things I learned during the past 21 years in school, during internships, and, above all, in university.

I would not have been able to write this thesis without the support of many people, to whom I am very grateful.

First of all, I thank Hans Ehm, Thomas Ponsignon, Stefan Degel, and Sebastian Eirich, for giving me the opportunity to write my thesis at Infineon Technologies on such an interesting topic, which made it also possible for me to stay in Munich (Germany) after my exchange semester. Furthermore, I am very grateful for the support of my company supervisor, Sebastian Eirich, and the endless time he took to support me with feedback. I could not have wished for better supervision. I also thank the Scenario & Flexibility Planning team of Stefan Degel for their warm welcome and the relaxing lunch talks and “digestion walks”. I always felt part of the team.

My gratitude also extends to Ahmad Al Hanbali and Marco Schutten of the University of Twente, for their helpful discussions and feedbacks to make this thesis more readable and to bring its quality to a higher level. Even though we were not able to meet every month personally, we still made it work through Skype or Lync.

Finally, I thank my family and friends who helped me get through the ups and downs during my entire studies, made sure I made time to relax from time to time, were always there for me when I needed them, and were understanding for the fact that I did not have that much time for them anymore during the last phase of my master thesis.

I wish you all a pleasant time reading this thesis.

Munich, 15th of September 2015.

Annemiek (Anna M.A.) Adriaansen

Management summary

Motivation

A semiconductor supply chain is complex and faces numerous challenges, such as volatile and uncertain demand, globalisation, a rapid changing environment, and long lead times. This results in demand- and supply-side uncertainties. Infineon hedges against these uncertainties by adding inventory and capacity buffers (also called contingencies). The Scenario and Flexibility Planning team built a supply chain simulation model, representing the internal supply chain, to study and balance these contingencies. However, the existing simulation model is not flexible enough to adjust to changes in the environment and is not yet accurate enough to ensure reliable results.

Research objective

The objective of this research is to develop a simulation-based framework to balance inventory and equipment contingencies for Infineon. For this purpose, we should provide the existing supply chain simulation model with flexibility and increase its accuracy (maximal deviation of 5% from the real data). With that model we want to know how to achieve a specific inventory service level at lowest total costs (inventory and capacity), and how this service level and the costs depend on stock levels and utilisation.

Supply chain simulation model

We extend the existing supply chain simulation model to improve its flexibility and accuracy. We achieve flexibility by enabling to load other bottleneck equipment into the simulation model automatically on model start-up. To increase its accuracy we improve the non-bottleneck delay determination and lot size calculation, and introduce factors that make up for data discrepancies. These improvements lead to the desired overall accuracy for each Key Performance Indicator (KPI), except for Layer Out per Week. We explain this by the fact that wafer losses are currently not included in the simulation model. Our major achieved accuracy improvement is for the cycle time, which improved from -21.9% to 1.1%.

Optimising the trade-off between inventory and equipment contingencies

We represent the inventory contingencies by the die bank and distribution centre (DC) stock levels and the equipment contingency by the utilisation and vary these in our experiments. We assume that the simulated optimal utilisation (in combination with the inventory contingency) corresponds with the fraction of the uptime that Infineon should use for the planning, also called the Plan Load Limit. For each simulated factor level combination we

collect the following KPIs: different service level types (adjusted non-stock out probability, fill rate, and adjusted fill rate) and the total costs. We use a simplified demand scenario as input for the experiments to be able to understand and quantify the relationships.

Balancing inventory and equipment contingencies

Based on the outcomes of the experiments we identify the following relationships:

- Lower costs do not always mean a lower service level.
- Producing at a low utilisation has a big negative impact on the costs although this positively influences the service level. For example producing at 80% instead of at 75% utilisation you can reduce total costs with about 5.5% (depending on the stock levels).
- Regardless of the service level type, a high inventory level is more important from a service level perspective, especially the inventory at stocking points close to the customer, because reaction time is the fastest there. DC stock is, however 2.6 times, more expensive than die bank stock in our case.
- For each service level type different factor level combinations lead to the lowest costs. To achieve certain adjusted non-stock out probability, both a high die bank stock level and a high DC stock level are important. To achieve a specific fill rate at lowest costs, having a high die bank stock level and utilisation is more important than having a high DC stock level. We explain this by the fact that the fill rate does not consider backorders of previous weeks. To achieve a certain adjusted fill rate at lowest cost, it is more important to have a high DC stock level instead of a high die bank stock level.

Recommendations

We recommend utilising the available equipment as high as possible and buffer at the DCs and die banks to make up for being less flexible and slower due to an increased flow factor (represents the variability in the process). What “as high as possible” means needs to be studied with a more detailed model. For further research, we propose Infineon to compare the current simplified model to a more detailed model. We laid the foundation by making the modelled bottlenecks flexible. Furthermore, we suggest to increase the reusability of the simulation model and to document the way data in the databases are obtained (when, where, and what) and to set-up databases that contain data specifically for simulation purposes. Moreover, we propose to include a more detailed backend and to extend the frontend object in the simulation model. For further experiments, we suggest to use a more realistic demand scenario with a broader product mix and to use a lower increment when increasing the factor levels. Moreover, we propose to replace the simplified planning function in the simulation model by a more advanced object.

Table of contents

Abbreviations

α	Variability
BE	Backend
CMOS	Complementary Metal–Oxide–Semiconductor
CT	Cycle time
DC	Distribution Centre
fab	Wafer fabrication facility
FE	Frontend
FF	Flow Factor
FIFO	First-In-First-Out
KPI	Key Performance Indicator
LOPW	Layer Out per Week
LSPW	Layer Starts Per Week
MTTF	Mean Time To Failure
MTTR	Mean Time To Repair
R&D	Research and Development
WIP	Work in Progress
SL^{α}	α Service Level: adjusted non-stock out probability
SL^{β}	β Service Level: fill rate
SL^{γ}	γ Service Level: adjusted fill rate
WOPW	Wafer Out per Week
WSPW	Wafer Starts Per Week

Symbols

α	variability
Δ_{tfk}	accuracy in week t for KPI k at facility f
λ_{TTF}	parameter for the exponential distribution for the time to failure
$bo_{ri,t-1}$	number of backorders for replication r product i at the end of the previous week t-1
bo_{rit}	number of backorders for replication r product i at the end of week t
cc	capacity costs
D_{avg}	average demand in units per day
$DelayNB$	delay of the non-bottleneck step
$D_{i\omega}$	demand until week t of product i
$d_{ri\omega}$	delivered products until week t of product i in replication r
FF	flow factor
FF_{hist}	historical flow factor
hc^{db}	holding costs at the die bank
hc^{dc}	holding costs at the distribution centres
PLL	plan load limit
$PR_{planned}$	planned productive time
$Qmin$	min quantile
$Qmin_{value}$	value of the min quantile
$Qref$	reference quantile

$\mathrm{Qref}_{\mathrm{value}}$	value of the reference quantile
RE	reach; the number of days demand can be fulfilled from stock when production would stop immediately
RPT	raw processing time
RTT	raw tool time based on 100 wafers
$\mathrm{RTT}_{\mathrm{lot}}$	raw tool time per lot
$\mathrm{scaleRPT}$	raw processing time scale
$\mathrm{SL}_{\mathrm{rit}}^{\gamma}$	adjusted fill rate for replication r product i and week t
$\mathrm{SL}_{\mathrm{rit}}^{\alpha}$	adjusted non-stock out probability for replication r product i and week t
$\mathrm{SL}_{\mathrm{rit}}^{\beta}$	fill rate for replication r product i and week t
ST	stock target
$\mathrm{s}_{\mathrm{rit}}^{\mathrm{db}}$	stock level at the die bank of product i at the end of week t in replication r
$\mathrm{s}_{\mathrm{rit}}^{\mathrm{dc}}$	stock level at the distribution centres of product i at the end of week t in replication r
TC	average total costs
$\mathrm{uniform}$	value drawn from the uniform distribution
UT	uptime as a share of the overall equipment time
$\mathrm{UT}_{\mathrm{planned}}$	planned uptime
UUm	uptime utilisation for manufacturing
$\mathrm{UUm}_{\mathrm{avg}}$	average utilisation
$\mathrm{UUm}_{\mathrm{hist}}$	historical utilisation

$\mathrm{UUm}_{\mathrm{planned}}$	planned uptime utilisation for manufacturing
$\mathrm{wc}^{\mathrm{BE}}$	Backend work in progress costs
$\mathrm{wc}^{\mathrm{FE}}$	Frontend work in progress costs
$\mathrm{WIP}_{\mathrm{rit}}^{\mathrm{BE}}$	work in progress in the Backend for product i at the end of week t in replication r
$\mathrm{WIP}_{\mathrm{rit}}^{\mathrm{FE}}$	work in progress in the Frontend for product i at the end of week t in replication r

1 Introduction

In the semiconductor industry, supply chain management faces a lot of challenges, such as volatile demand, globalisation, and long lead times. Therefore, companies operating in this industry are looking for ways to deal with these challenges. Infineon Technologies AG (Infineon), with headquarters located in Neubiberg, near Munich, is such a company. To help their Supply Chain Operations Scenario and Flexibility Planning team by improving their simulation model of the whole internal supply chain and providing use-cases of the simulation model, we conduct this research. The research is part of the Master study "Industrial Engineering and Management", with a focus on production and logistics management.

In this chapter, we outline this project. Section 1.1 gives a brief description of the organisation we conduct this research for. Section 1.2 and Section 1.3 describe the motivation for this quantitative research and the research problem, respectively. Section 1.4 states the objective of the project and Section 1.5 scopes the project. The last section of this chapter, Section 1.6, provides the research approach.

1.1 Organisation

In Section 1.1.1 we briefly introduce the company that the research is conducted for, Infineon and in Section 1.1.2 we describe its supply chain.

1.1.1 Infineon Technologies

Infineon is a leading semiconductor manufacturer with a focus on three application areas: energy efficiency, mobility, and security. For these applications, Infineon produces tens of thousands of different products. These products are divided into two technology classes: power semiconductors and complementary metal–oxide–semiconductors (CMOS). The power technology class focuses on high current and low resistance semiconductors, and the CMOS class on high switching frequency and high density semiconductors. Power technology is the core business of the company. (Infineon Technologies AG, 2014)

The corporate supply chain department of Infineon is responsible for connecting the global value chain. At Infineon, supply chain management is regarded as a competitive advantage.

1.1.2 Infineon's supply chain

The Infineon supply chain consists of different stages. Figure 1-1 visualises a simplified supply chain as typical for a semiconductor manufacturer. The customers (and customers' customers) and suppliers (and suppliers' suppliers) are not included in this figure, because they are not in the focus of this research.

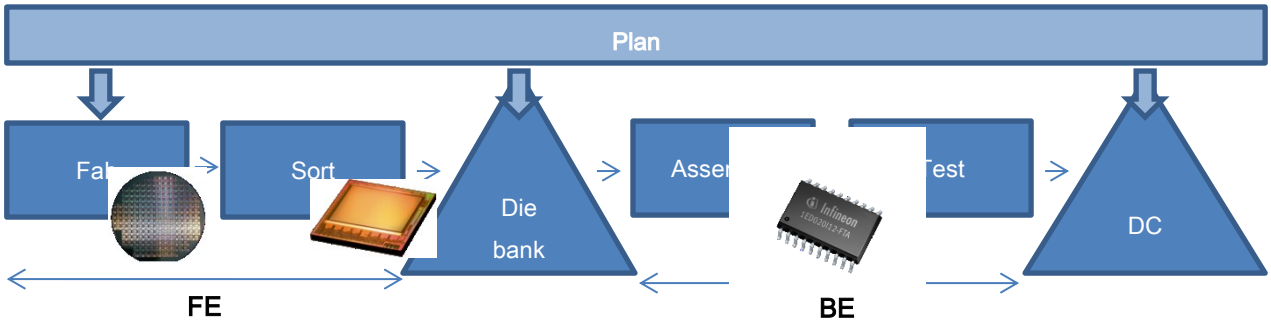


Figure 1-1 Simplified supply chain visualisation of Infineon Technologies

To provide insight in the way Infineon has designed its supply chain, we give a simplified description of the flow through the *manufacturing levels* (encompassing fab, sort, assembly, and test) and stocking points, as shown in Figure 1-1, focusing on the components relevant for this research:

1. The wafers are treated in a wafer fabrication facility (fab). This treatment consists generally of layering, etching, doping, polishing, cleaning, and lithography, but occasionally, for a few products, some of the treatments are left out. The sequence and the frequency of these steps are product type dependent. As these production steps regularly need to be repeated, the equipment for these steps is often visited multiple times by the same wafer. This results in up to 500 steps for the treatment of one wafer.
2. The wafers go to the so-called 'sort' manufacturing level after their treatment in the wafer fab. Here the produced dies on the wafers are tested and the bad dies on the wafers are marked. One wafer consists of dies of one type.
3. The dies are stored in an intermediate inventory: the die bank. There are thousands of different die types.
4. At 'assembly', the wafers are sawn and the chips are assembled. This step consists of wire bonding, die bonding, moulding, and/or trim and form. This is, again, product type dependent.
5. In the 'test' manufacturing level, the chips are tested and bad chips thrown away.
6. The finished products are stored in a distribution centre (DC).

The first two manufacturing levels form the front end (FE) of the production process. Most of the value of the end-product is added in these stages. The FE *cycle time* (CT), defined as the length of time spent by a product unit in the FE system, from the release of the wafer into the fab until finishing the last step in sort, ranges typically from 40 to 100 days. In the FE, a Make-to-Stock strategy is typically followed, because the demand is often not yet known when production needs to be started to be able to deliver on time to the customers with short *lead times*. The lead time is the delivery time communicated to the customer. (Ehm, 2015)

Manufacturing levels assembly and test form the back end (BE) of the production process. The BE CT is shorter than the FE CT and typically spans from 5 to 20 days. The BE CT is defined as the length of time spent by a product in the BE system; from the release of the semi-finished product in assembly until the product completes the testing step. In the BE an Assemble-to-Order strategy is typically followed. (Ehm, 2015)

Infineon typically works with process groups in the FE and with packages in the BE. Process groups summarise products with similar process flows. A package consists of chips that are assembled the same way.

Figure 1-2 Locations of FE and BE facilities of Infineon Technologies

At Infineon, the different steps in the supply chain are performed in different cities and countries (Figure 1-2). In this figure (and this thesis) the integration of International Rectifier, a company taken over by Infineon in 2015, is not included. The FE facilities are located in Villach (Austria), Kulim (Malaysia), Dresden (Germany), and Regensburg (Germany). The die banks are located both at FE and BE facilities. The BE facilities are located in Malacca (Malaysia), Regensburg (Germany), Warstein (Germany), Batam (Indonesia), Wuxi (China),

Beijing (China), Cegléd (Hungary), Singapore (Indonesia), and Morgan Hill (USA). The DCs are located in Asia, Europe, the United States of America, and China.

In the supply chain there are different ways to *buffer*. We define buffers as an excess resource that corrects for misaligned demand (e.g. due to forecast errors) and transformation (e.g. unforeseen losses), and takes on one of the three forms: inventory, time, and capacity (Hopp & Spearman, 2008). To be in-line with the company terminology, we refer to buffers as *contingencies*. In this thesis, we focus on inventory and capacity contingencies: the intented and planned spare capacity or inventory, which can be used to rapidly hedge against uncertainties. To improve the alignments of these contingencies, Infineon started to build a simulation model in 2014. We refer to this model as the "existing model". To the version that we develop for this thesis we refer as the "updated model".

The objective of this research is to find a financial benificial balance between inventory and equipment contingencies under service level requirements by the use of discrete event simulation. Before being able to determine the balance between the considered contingencies, the existing simulation model has to be improved with regard to accuracy and flexibility. These contingencies have to ensure that a certain service level type at the DC is at a sufficiently high level. We explain which service level types we take into account in Section 1.2.

1.2 Research motivation

Supply Chain Management in the semiconductor industry is a challenge, because of the rapidly changing environment. Since the transistor was invented in 1948, the number of transistors on a chip is approximately doubling every two years at decreasing costs per megabyte; this is also known as Moore's law. The number of applications has increased over the years, which has also increased the demand and competition in the industry (Gupta, Ruiz, Fowler, & Mason, 2006). In the past few years it became easier, but also necessary, for companies to expand or move processes to other continents. This is not only considered for financial benefits, such as cheaper labour, but also to be closer to large customers. This globalisation does not only make Supply Chain Management more important, but also more challenging (Ehm, Ponsignon, & Kaufmann, 2011; Fowler & Rose, 2004; Jain, Lim, Gan, & Low, 1999). Furthermore, the industry faces volatile demand and the products have short life cycles and steep product ramps (Ehm et al., 2011; Brown, Lee, & Petrakian, 2000). In addition, demand is uncertain. This is because production needs to be started in anticipation of future demand to achieve competitive lead times for the customers.

Next to demand uncertainty, supply-side (production) uncertainty exists. Supply-side uncertainty is caused by all factors that contribute to uncertain future output quantities. For example, unstable CTs, together with yield losses of the (semi-)finished products, cause supply-side uncertainty.

In the semiconductor industry the wafer fabs are the most complex and expensive facilities. This complexity is caused by the numerous and time-consuming process steps that are product specific and performed on expensive equipment (Fowler & Rose, 2004; Gupta et al., 2007; Uszoy, Lee, & Martin-Vega, 1992). Most of this equipment is provided with very advanced technology and operators have to deal with unpredictable equipment failures or preventative maintenance (Gupta et al., 2006; Uszoy et al., 1992). This further increases the CT variability. Consequently, CTs are unstable, which leads to longer lead time commitments to the customers (Gupta et al., 2006).

In this section we explained causes of both demand- and supply-side uncertainties that exist in the semiconductor industry. These uncertainties make it difficult to achieve a high service level which is essential in the highly competitive semiconductor industry. Therefore, semiconductor manufacturers need to find a way to cope with these uncertainties. One way to do so is to add inventory buffers of finished products, using product postponement and/or process postponement. Another way is to add spare capacity to equipment. Both are, however, associated with high costs.

For the service level we consider three types for this research, because they all provide different information about the performance of the supply chain and the relating customer service. These *service level types* are: *α, β, and γ* service level (SL^{α}, SL^{β}, and SL^{γ} respectively). We define the SL^{α} as the probability that demand can be met completely during a week, without considering backorders of previous weeks; this can be either 0 (not all demand can be met) or 1 (all demand can be met). This is an adjusted version of the non-stock out probability known in the literature in which backorders of previous weeks are considered. The SL^{β} provides the proportion of order quantities fulfilled by stock including backorders of the current period and is also known in literature as the fill rate (Axsäter, 2006; Minner, 2012). The γ service level indicates how fast a production system can recover from all backorders and is also known as the adjusted fill rate or ready rate (Axsäter, 2006; Minner, 2012).

1.3 Problem description

As explained above, semiconductor supply chains face various types of uncertainties. Therefore, semiconductor manufacturers add contingencies to hedge against these uncertainties. Typical kinds of contingencies are semi-finished and finished goods inventories, excess capacity, and inflated demand assumptions. However, the rapid price declines lead to lower margins. On the other side, lithography equipment costs several million euros and makes capacity unattractive. Moreover, interdependency between equipment *uptime utilisation for manufacturing* and inventory levels exists. We define the uptime utilisation for manufacturing as the ratio of time equipment is actually producing to the time the equipment is available for production (see Section 2.1.2). For convenience we will refer to the uptime utilisation for manufacturing as utilisation from now on. To explain the interdependency between equipment utilisation and inventory levels we give an example: running a wafer fab at a high utilisation to keep the return on investment high increases the manufacturing CT, which might require increased inventory levels to stay reactive towards demand variations on short notice, and to hedge against lost production time, because of equipment failures and more frequent maintenance measures that are needed at higher utilisation (Slack, Chambers, & Johnston, 2007). Therefore, Infineon aims at finding a balance between inventory and equipment contingencies, so that they are able to deliver on time to their customers at minimal costs. We expect that by aligning these contingencies competitive advantages can be gained.

For this purpose, Infineon started to set up a simulation model that is suitable to study the phenomena described above. Since the system that is being modelled is complex, non-linear, and dynamic, and has to cope with several uncertainties and a large number of parameters, the development of a simulation model is preferred to analytical modelling (Borshchev, 2013). Nevertheless, supply chain simulation is, besides the challenges of the semiconductor industry, a challenging, effortful task itself. This effort consists of high time to build, run, and maintain such a model. To cope with this problem, several approaches exist. An example, as Infineon is doing already, is building a model with a reusable library (Yuan & Ponsignon, 2014). Another example is increasing the flexibility of the model, which makes it easier to build future or extend existing simulation models. In this way, the model can be (partly) reused, and the time effort to build and maintain a model to answer questions in the future is reduced. Another approach is reducing the level of detail and to keep the model as simple as possible (Brooks & Tobias, 2000). The challenge here is to not sacrifice the required accuracy for the targeted question.

The Scenario & Flexibility Planning team at Infineon is currently building such a model on the internal supply chain level, but improvements can be made regarding the flexibility and accuracy of the model. For example, the model has to be flexible enough so that changes in the manufacturing environment, such as additional facilities or a different number of bottleneck work centres, and other scenarios, such as additional contingencies, can be incorporated easily in future studies. Therefore, this research identifies and implements possible improvements for the flexibility in the wafer fab part of the model. The accuracy and of the model is determined by comparing the simulation data with real data (see Section 2.2.4 for more details).

1.4 Research objective

Based on the problem description we formulate the objective of this research. The objective of this research is to develop a simulation-based framework to balance inventory and equipment contingencies for Infineon, and by improving the existing supply chain simulation model with regard to flexibility and accuracy. In essence, we want to know how to achieve a specific service level at lowest costs. The improved simulation model should give a valid representation of the real world; we aim for increasing the overall accuracy of the simulated data to a maximal deviation of 5% from the real data.

1.5 Scope

The wafer fab is the most critical facility from a CT, cost, and value perspective. Therefore this research focuses on improving this part of the existing supply chain simulation model. Since there is already a preliminary simulation model of the wafer fab, a lot of existing data is available for this research. However, we need to check if this data fits our own simulation purposes. The experiments regarding balancing the equipment of the wafer fab and inventory contingencies are conducted with the whole internal supply chain simulation model, including all manufacturing levels in the supply chain. For this purpose each manufacturing level is modelled as an object. The wafer fab object in detail and the other objects less detailed. Moreover, we need to define a demand scenario that represents the customer order input. Figure 1-3 graphically shows this delineation of the project.

Furthermore, this thesis focuses on equipment contingencies of the bottlenecks in the wafer fabs (i.e. not in sort, assembly, and test) and inventory contingencies of the die bank and DC level, to cope with the previously mentioned uncertainties. Business contingencies, for instance, are not taken into account. Nevertheless, we need to keep in mind that these other contingencies have to be integrated in the simulation model for future studies.

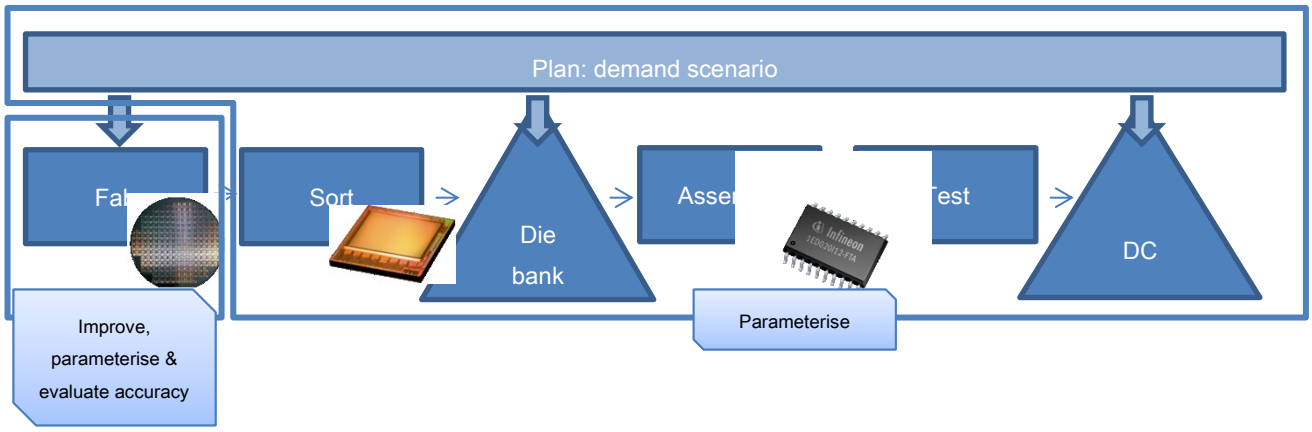


Figure 1-3 Simulation model scope

Considering the limited time available for this project, the system of which we want to build a simulation model needs to be simplified. Only wafers with a diameter of 200 mm (8") are taken into account, because these are processed most. Consequently, the 100 mm (4"), 150 mm (6"), and 300 mm (12") wafers are left out. These wafers are in most cases processed on dedicated equipment and therefore do hardly overlap with the 8" equipment. Moreover, only in-house facilities are considered, i.e. external foundry partners and subcontractors are left out. Furthermore, this research only focuses on the products with the power technology class of Infineon, because this is the company's core business. This also means that the 200mm fab in Dresden is left out, because they mainly process wafers with the CMOS technology class.

The products produced on 8" wafers by Infineon's own facilities belonging to the power technology class represent about 50% of Infineon's production volume.

1.6 Research approach

In literature, different authors have formulated the steps that need to be taken in a simulation study (e.g., Banks, 1998; Fowler & Rose, 2004; Law, 2007; Shannon, 1988). The steps are sometimes named differently and are sometimes less or more aggregated compared to each other, but generally they come down to a similar sequence of actions. As there hardly exists a difference between the steps in the different literature, we decide to take the steps from the well-known book of Law (2007).

Law (2007) defined ten steps in a simulation study (Figure 1-4). We add the possibility of new information that becomes available at the validation step, so that the previous steps are revised in case of new relevant information. We defined the research questions using this step-by-step approach of Law (2007). This chapter covers the first step.

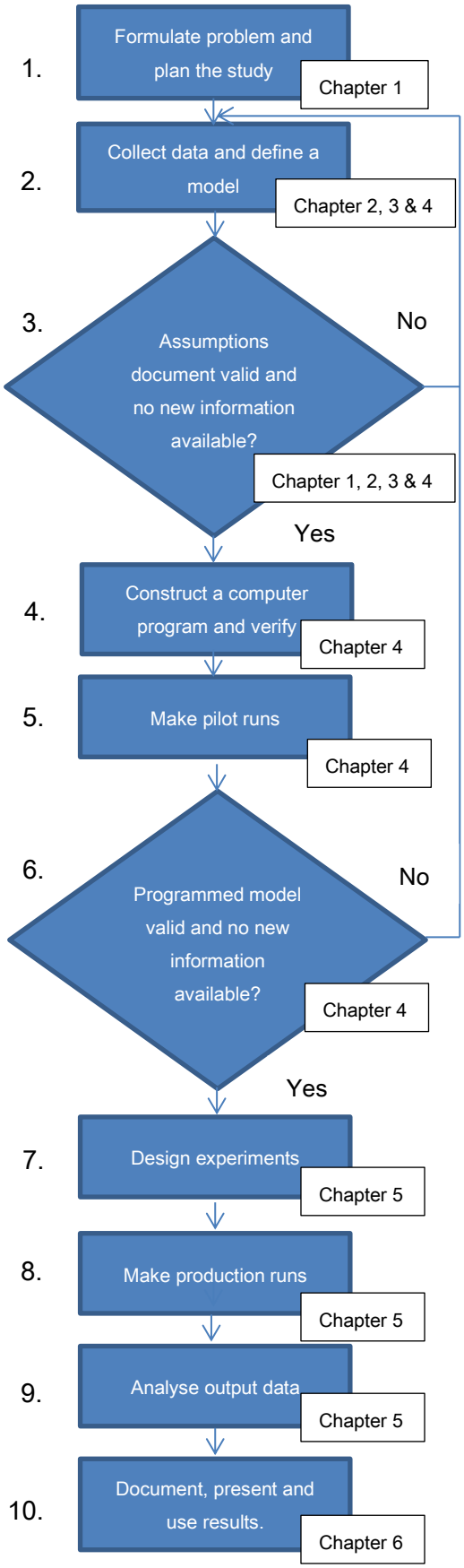


Figure 1-4 Ten steps in a simulation study, adapted from Law (2007, p.67)

To solve the problem described in Section 1.3 and to reach the goals as described in Section 1.4 the main research question that we need to answer is:

How can the equipment and inventory contingencies be balanced in a most cost efficient way for a defined service level using a flexible and accurate supply chain simulation model of Infineon Technologies?

To answer this main question, the following questions are set up. By answering these sub-questions, we are able to give a solid answer to the main question.

1. What is the current status of the supply chain simulation model and how are the inventory and equipment contingencies planned currently?

We answer this first sub-question by examining and describing the current situation of both the simulation model and the current way of planning inventory and equipment contingencies in Chapter 2. This includes the identification of the most important performance measures. We discuss these topics during meetings with experts of the company to gain professional insights as well. This step is necessary to be able to compare the findings in literature with the current situation, so that we can identify possible improvement areas (see step 2 in Figure 1-4). Furthermore, we will be able to compare the results of the experiments with the current situation.

2. What literature is available related to the topic of balancing equipment and inventory contingencies?

To answer sub-question 2, we conduct a literature study on the topic of equipment and inventory contingencies as specified by step 2 (Chapter 3). In this way, we can design sound experiments.

3. Based on the existing literature, what should be taken into account when developing a supply chain simulation model and in what ways can the wafer fabrication facility part be modelled accurately and flexibly?

To answer sub-question 3, we conduct a literature study on supply chain simulation and on the modelling of wafer fabrication facilities. This also belongs to step 2 and we discuss this in Chapter 3 as well.

Step 3 is about *validating* the collected data and the conceptual model. Validation means that we have to check if the conceptual model is an accurate representation of the real world. We do this during steps 1 and 2 and experts will do a final check in step 3.

Based on the outcomes of the first three sub-questions we improve the simulation model in step 4. We describe the conceptual model and its implementation in the existing supply chain simulation model in Chapter 4.

4. How can the supply chain simulation model be verified and validated?

After we have implemented the improvements, we *verify* and validate the simulation model to make sure that the model is accurate. This means solving all the bugs in the simulation model, so the simulation model can run smoothly, without errors, and checking that the simulation model matches the conceptual model. This corresponds to steps 4, 5, and 6 and we describe this in Chapters 3 and 4.

5. How should the supply chain simulation model be parameterised to be able to determine the desired balance for inventory and equipment contingencies?

Once the simulation model is valid, we design experiments to find a balance between inventory and equipment contingencies by varying parameters that represent these contingencies. This belongs to step 7 and we describe this parameterisation in Chapter 5. Furthermore, we need to define a demand scenario to be used as input of the simulation model. The answer on the sub-question 5 gives the input to the experimental design.

6. How sensitive is the supply chain simulation model to changes in the values of the parameters?

A sensitivity analysis is included in the experimental design in Chapter 5 by varying the parameters defined in the experimental design. After the execution of the experiments, step 8, we answer this sub-question 6.

7. How can the equipment and inventory contingency plans be balanced taking costs and the service level types into account?

We answer the last sub-question based on the data analysis of the experiments in Chapter 6 as specified by step 9 and this sub-question completes the answer to the main research question. We draw a conclusion in Chapter 6, in which we also present the implications. This corresponds with the last step in a simulation study, step 10.

2 Analysis of current situation

Section 2.1 describes the relevant aspects of the inventory and capacity management of Infineon. The existing supply chain simulation model, with a focus on the wafer fab part, is described and analysed in Section 2.2. Finally, Section 2.3 draws conclusions based on the analysis in the previous sections.

2.1 Inventory and equipment contingencies

By the use of contingencies, Infineon hedges against demand and supply (production) uncertainty.

Section 2.1.1 describes the inventory management practices, including inventory contingencies. The capacity management practices of Infineon, including equipment contingencies, are outlined in Section 2.1.2.

2.1.1 Inventory Management

The *inventory contingency* is meant for buffering fluctuations in the production network and order behaviour, as well as acting as a source for fast delivery by using (semi-)finished goods. The buffering is mainly done at the die bank and the DCs.

At Infineon the supply chain planner is responsible for setting stock targets for different stock types, such as safety stock or ramp-up stock. This target setting is done based on a set of rules with the help of a software tool. The supply chain planner is able to manually set the rules in the tool, which is mostly based on his experience.

The stock target (ST) is the quantity of stock that is needed to cover "*reach*" days of expected demand for each of the stock types. Hence, the reach is the number of days the supply chain planner wants the inventory to last, to be able to fulfil the average daily demand if production would stop immediately. The stock target is calculated by multiplying the average demand in units per day (D_{avg}) with the reach (RE) (Equation 1).

$$ST = D_{avg} * RE \quad (1)$$

The supply chain planner has to set six other parameters, next to the reach: the stock target per product, the manufacturing level, so where he wants to place the stock (for example, at a DC or die bank), the stock type, and the demand type, such orders and/or forecasts. Furthermore, besides the demand type he has to choose between using the constrained or unconstrained demand and which periods he wants to consider for the average demand. We

explain these other six parameters in more detail in Appendix A, because they are not relevant for this thesis; only the reach is important for the inventory contingency.

The reach is product and situation dependent and its value is mainly based on the experience of the planner. It is defined as a multiple of seven in most cases. A reach of 28 days at both the die bank and the DC is most commonly used by the supply chain planner for the inventory planning at Infineon.

2.1.2 Capacity Management

Equipment contingency is meant for buffering against WIP waves, equipment breakdowns, and demand uncertainties. This buffer consists of planned spare production capacity time, also called the *standby time*, which can be used in reality if needed. The standby time is defined as the time or the share of the overall equipment time the equipment is in a condition to perform its intended function, but there is no operator, product or support tool available (Oechsner, et al., 2003).

The standby time is closely interlinked with two concepts that are of significance for this thesis: *Overall Equipment Effectiveness* and *Operating Curve Management*. Overall Equipment Effectiveness helps to understand the terms and Operating Curve Management to understand the consequences. This section explains these concepts and describes how the capacity buffer is subsequently planned.

Overall Equipment Effectiveness

The standby time is one of the machine states used by the Overall Equipment Effectiveness measurement concept and is often expressed as a percentage of the overall equipment time. The Overall Equipment Effectiveness is measured at Infineon by taking the product of the proportions of the availability of a machine (also called availability efficiency), the utilisation (also called uptime or operational efficiency), the process performance (also called rate efficiency), and the yield (also called quality efficiency) (Oechsner, et al., 2003; Infineon Technologies, 1998; Slack, Chambers, & Johnston, 2007).

To determine the Overall Equipment Effectiveness, Infineon measures ten machine states (Infineon Technologies, 1998). To gain a clear understanding of this concept, we aggregated and reduced the machine states to the following four components (for detailed information see Appendix B): downtime, non-sales production time, sales production time, and standby time.

The non-sales (R&D) production time plus the sales production time is the time equipment is producing units and is also called the *productive time*. When the productive time is expressed as a share of the *uptime* (productive time plus the standby time), it represents the utilisation (Equation 17, Appendix C).

These components are usually presented in a graph as for example shown in Figure 2-1. The graph shows how the overall equipment time is distributed over the components. The overall equipment time is a selected time period, for instance 168 hours, in which the equipment is present. In this time, the equipment can be used or not used. The downtime of the equipment corresponds to 12% of the overall equipment time, so the equipment was down for about 20 hours. The equipment was used for almost 98 hours for sales production, about 25 hours for non-sales production and was standby for about 25 hours as well.

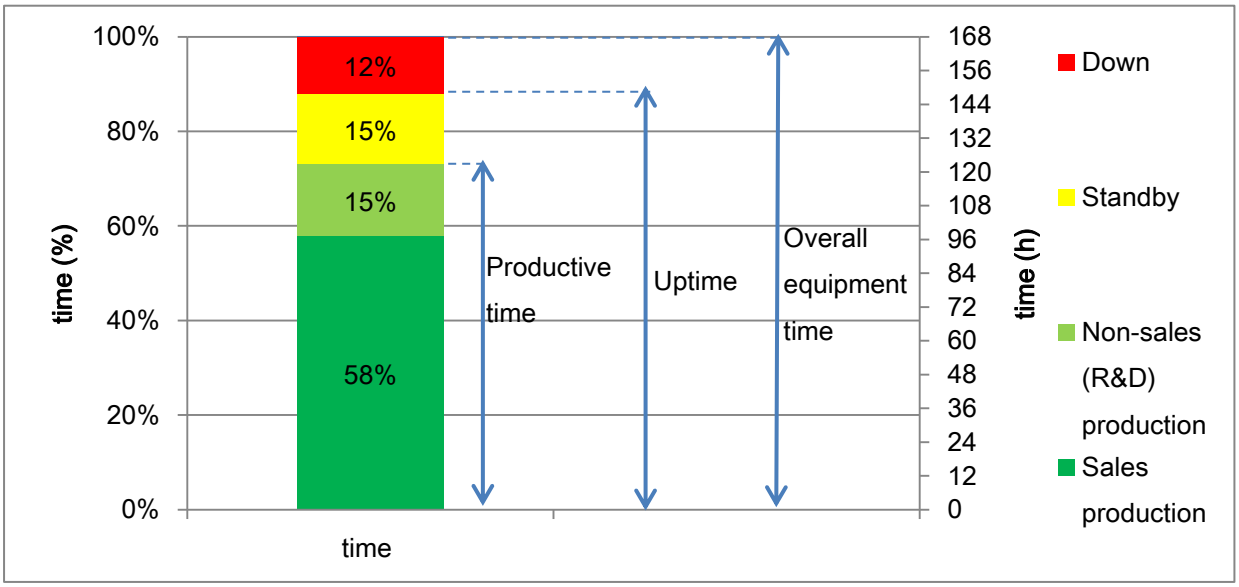


Figure 2-1 Example of an Overall Equipment Effectiveness graph

Operating Curve Management

Operating Curve Management is a methodology based on queuing theory (Aurand & Miller, 1997). This sub-section describes the Operating Curve concept globally. A more detailed description, including the link to queueing theory, can be found in Appendix C.

The planned standby time and the planned productive time of the equipment are an input for capacity planning. The amount of planned standby time influences the *Flow Factor* (FF), which represents how much the realised CT is bigger than the *Raw Processing Time*. The Raw Processing Time is the planned average cycle time needed for the process to meet the final performance criteria for the product in case of optimal conditions based on the average

lot size. It excludes the waiting time and process inefficiencies (Hopp & Spearman, 2008; Infineon Technologies, 2010). When no standby time is planned, the planned utilisation of the equipment is 100%, which leads according to Operating Curve theory to an infinitely high FF, thus to infinitely long CTs (see Appendix C).

The operating curve visualises the relation between the FF and the utilisation (see Figure A-1 in Appendix C). It shows the trade-off between high speed (low FF) and high throughput (high utilisation) when resources are limited. Operating Curve Theory also takes the *variability* (α) into account. α describes the quality of the non-uniformity of attributes in manufacturing systems, such as products and processes. Several characteristics, such as process times, machine failures, and quality measures, are prone to this non-uniformity (Hopp & Spearman, 2008). A low α indicates a good line performance. α is calculated by using the FF and the utilisation (UUm):

$$\alpha = \frac{(FF - 1) * (1 - UUm)}{UUm} \tag{2}$$

As stated above, the FF will be higher in case of high utilisation of equipment (see Operating Curve theory in Appendix C). This typically also negatively affects α. The operating curve is used by Infineon as an indicator of the production process performance.

Capacity planning

As described before, running a wafer fab at 100% utilisation is not desired. Therefore a *Plan Load Limit* is needed. This limit determines the maximal planned utilisation and therefore the maximal productive time.

The capacity planner at Infineon is responsible for the capacity planning in the FE. This planner sets the planned *uptime*, which is the time the equipment is available for production. This consists of the standby time and the productive time. The remaining time is the planned downtime. Furthermore, the planner determines the Plan Load Limit of the planned uptime. The Plan Load Limit is the planned productive time of the equipment and is an empirically developed value based on historical data, tool stability, number of equipment and most of all, experience. This Plan Load Limit is set for an undetermined period, and only changes when there are changes in the system, for instance if the number of equipment alters. The Plan Load Limit typically varies between 94% and 75% of the uptime. Consequently, the planned utilisation ($UUm_{planned}$) is at most equal to the Plan Load Limit (PLL), which equals the maximum of the planned productive time ($PR_{planned}$) as a share of the planned uptime ($UT_{planned}$):

$$\max. UUm_{planned}\,(\%) = PLL\,(\%) = \frac{\max. PR_{planned}}{UT_{planned}} \qquad (3)$$

The Plan Load Limit may be exceeded in the actual production, but on average this limit should be maintained to secure the production flow. For a better understanding of the important terms above, they are graphically represented in Figure 2-2.

To give an example, for a single tool, if the overall equipment time is 168 hours of which the share of the uptime is 75%, the uptime is 126 hours. If the Plan Load Limit is then set at 94% of the uptime, this means that 118.44 hours capacity is available for the planning system (see Figure 2-2). The Plan Load Limit limits the planned production volume. When the Plan Load Limit is set high, the planned utilisation can be high. Subsequently, the planned CT becomes long, the uncertainty increases and the inventory levels should be high as well to fulfil demand. On the other hand, a low Plan Load Limit and therefore a low planned utilisation leads to a better product flow, because, of low CTs and less uncertainty. However, this requires more capacity to fulfil the same demand, which is expensive and therefore not always a preferred option. This is the trade-off we want to analyse in this research.

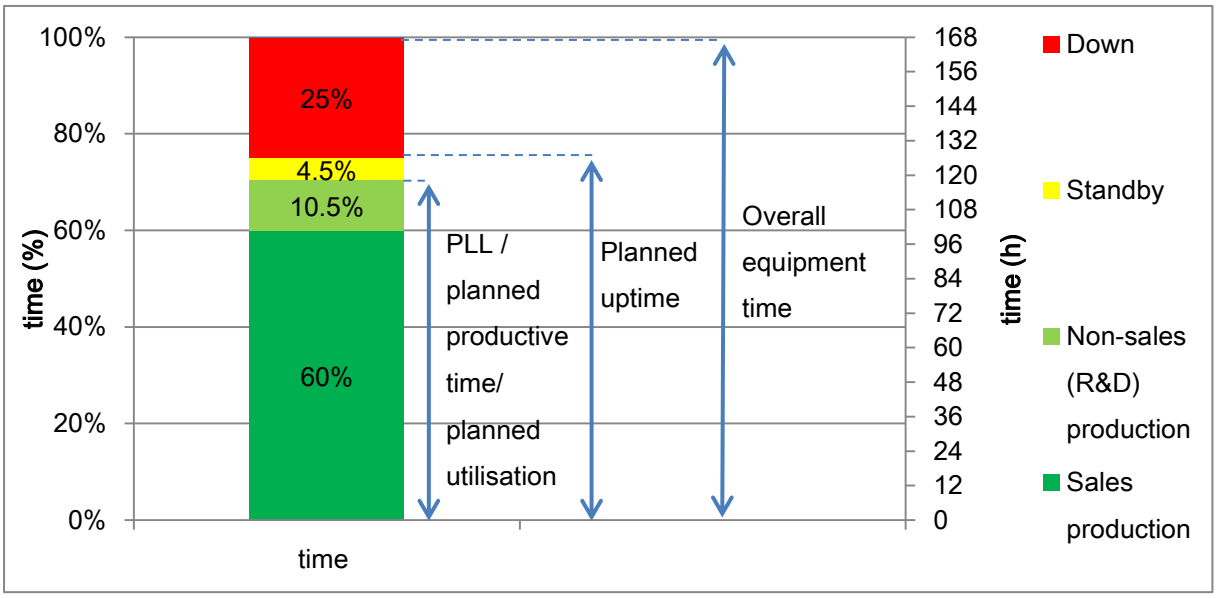


Figure 2-2 Equipment planning terms

By varying the Plan Load Limit in this research and balancing this limit with the stock reach, the best fit of the inventory and equipment contingencies can be determined. We use discrete-event simulation to determine this best fit. This method is chosen, among others, because Infineon's supply chain is a complex system (this is reasoned more extensively in Section 3.2).

2.2 Supply chain simulation model

Infineon started to build a supply chain simulation model to be able to study contingencies. In this section, we introduce the existing version of the simulation model (Section 2.2.1). Section 2.2.2 describes the wafer fab part of this supply chain simulation model. Furthermore, Section 2.2.3 describes the current inputs and outputs, followed by Section 2.2.4, which indicates the accuracy status of the simulation model. Section 2.2.5 discusses possible problems and improvement areas of the simulation model.

2.2.1 Supply chain simulation model for contingency optimisation

The Supply Chain Innovation department of Infineon has defined four different levels for supply chain simulation models, presented in order of increasing scope:

- the *work centre level*, which represents tools,
- the *factory level*, in which a single facility is modelled,
- the *Infineon supply chain level*, in which the internal supply chain is modelled,
- the *end-to-end supply chain level*, which is the internal and external supply chain. This level includes silicon foundries, subcontractors, direct customers, end customers, and suppliers.

The last two levels are suited best for scenario analysis, to gain insight into the impact on the supply chain.

The supply chain simulation model for contingency optimisation covers the third level and is implemented in AnyLogic professional edition version 7.0.3 (AnyLogic, 2015). It is a discrete-event simulation model. Law (2007, p.6) defines discrete-event simulation as “the modelling of a system as it evolves over time by a representation in which the state variables change instantaneously at separate points in time”. These points in time are dependent on specified events in the simulation model. Moreover, an object-oriented approach is used for the simulation, in which objects interact with each other over time (Law, 2007). The model is schematically shown in Figure 2-3. It covers three main areas: data, plan, and make.

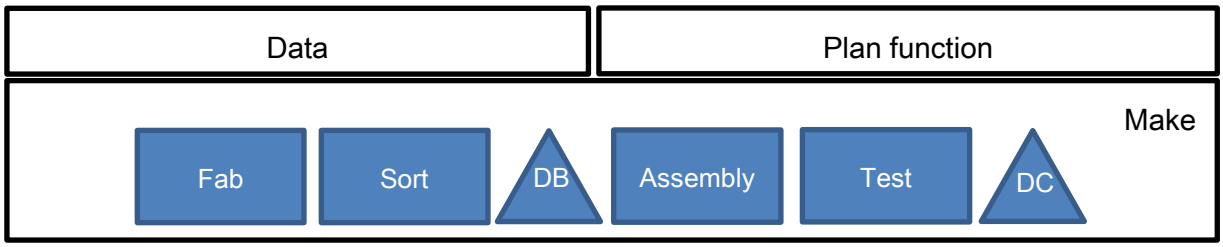


Figure 2-3 Top layer of the supply chain simulation model

The plan function, as well as the data, and the objects of the manufacturing levels are represented at the top layer of the model (Figure 2-3). The plan function is called in the simulation model at the beginning of each simulated week. The main purpose of the plan function is to calculate how many lots of which product type need to be released during the following week, considering WIP, actual and targeted inventory levels, orders, forecasts, and unfulfilled requests from the previous weeks. We provide and explain the used formulas in Appendix D. We note that the current plan function incorporates no production smoothing, such as scheduling the demand earlier for production if capacity in a later period is not sufficient.

The required data is provided by the data layer (see Section 2.2.3). Once a lot is released into the make part, the lots are processed according to the routes, specified in the data layer. In the make part all relevant manufacturing levels (fab, sort, assembly, and test) in the supply chain are modelled. As the wafer fab object is the focus for improvements, we explain this object in more detail in Section 2.2.2.

2.2.2 Wafer fabrication facility simulation model

The current wafer fab part of the simulation model is modelled at an intermediate level of detail and is a simplified representation of a semiconductor wafer fabrication facility. Brooks and Tobias (2000, p.1010) define this simplification as “a reduction in the number of components or connections”. This can for instance be done by removing components, such as non-bottleneck machines, or by aggregation, such as grouping the different products into product families (Brooks & Tobias, 2000).

The simulation model consists of five layers, as shown in Figure 2-4. In the first, main, layer, all manufacturing levels are represented. Here all inputs and desired outputs are defined. We discuss these in Section 2.2.3. The fab object consists of different FE facility objects. In the simulation model, there are three FE facilities modelled representing Villach, Regensburg, and Kulim. These are the facilities where the 8” wafers used in the power technology are produced (see scope in Section 1.5). There is a delay object called ‘Others‘, which represents all other FE facilities, such as the facility in Dresden and the partner foundries. This delay time is specified by the planned CT, which is defined in the master data. The Villach, Regensburg, and Kulim objects contain instances of the production unit object. This is only shown for Villach in Figure 2-4. Note that the next layers, layer 3, 4, and 5, are also embedded in the other FE facility objects. At the production unit layer, some inputs that are important for the parameterisation of the equipment in the next layer are gathered. In this

next layer the equipment is modelled as stepper (lithography) work centres, sputter work centres, and non-bottleneck work centres. The non-bottleneck equipment is considered in the simulation model as a single delay (5b). The stepper and the sputter work centres are modelled in more detail (5a).

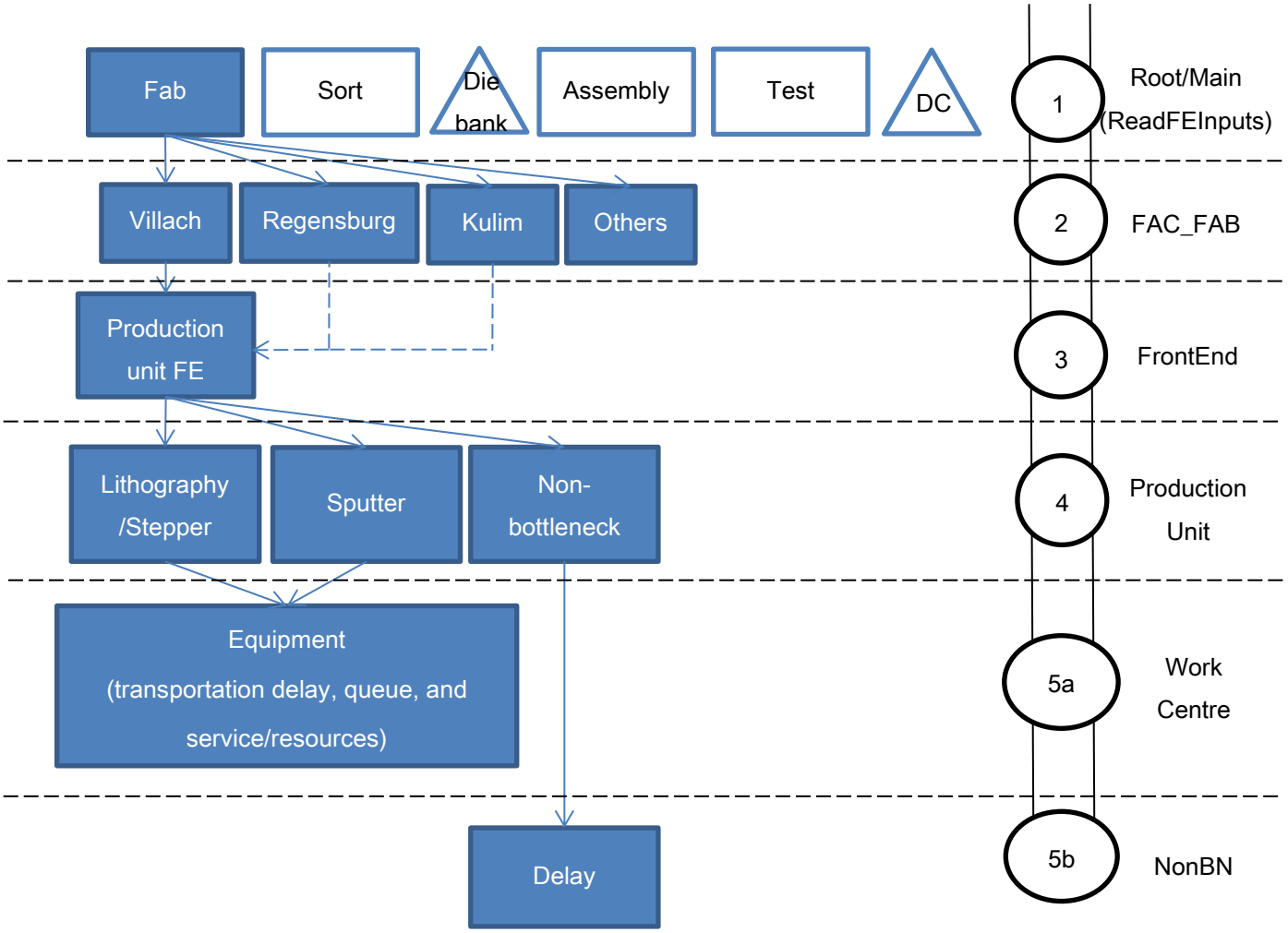


Figure 2-4 Wafer fabrication facility model in the existing simulation model

Both bottleneck work centres have a certain number of tools assigned with a specific mean time to failure (MTTF) and mean time to repair (MTTR). The TTF and the TTR are both assumed to be exponentially distributed in the simulation model with parameter λ. For the TTR a mean (μ) of 1.25 hour is used (where 1/μ=λ). We calculate the λ for the TTF (λ_{TTF}) in hours^{-1} using the uptime (UT) as a fraction of the overall equipment time. This is available from historical data (see Section 2.2.3):

$$\lambda_{TTF} = \frac{1}{\frac{UT}{1-UT} * 1.25\ hour} \quad (4)$$

The lots in the queue in front of the bottleneck tools are prioritised according to *first-in-first-out* (FIFO) in the simulation model. However, in reality this is done according to the *Earliest Operation Due Date*. This means that typically a lot with the earliest due date for the pending operation can leave the queue first (Smith, Minor, & Jen, 1995). However, in the simulation model the first lot arriving at the queue can leave the queue first.

Non-bottlenecks are modelled in less detail than the bottlenecks. They are represented by a single delay object. The delay time is determined based on the Operating Curve Theory and is approximated by a theoretical FF measured at the end of the FE process (when leaving the wafer fab), and the Raw Processing Time. To do so, first α is calculated outside the simulation model by using the FF and the utilisation (UUm) by Equation 2 in Section 2.1.2. The utilisation and the FF used for this calculation are derived from historical data.

Afterwards the FF is estimated in the wafer fab simulation model per location using the α, that was calculated outside the model, and the simulated average of the utilisation of the stepper equipment (UUm) (they assume this equipment to be most critical), measured at the time the lot arrives:

$$FF = 1 + \alpha * \frac{UUm}{1 - UUm} \tag{5}$$

The delay of the non-bottleneck equipment (DelayNB) is calculated by multiplying the Raw Processing Time (RPT) with this FF (Equation 6, derived from Equation 17 in Appendix C). This is implemented as an estimate of the CT.

$$DelayNB = RPT * FF \tag{6}$$

The Raw Processing Time (RPT) is determined by multiplying the Raw Tool Time (RTT) with a predefined factor, called scaleRPT in the simulation model. This scaleRPT is currently empirically estimated by the developers of the model and makes up for the transportation time between equipment, which is not included in the raw tool time. The Raw Tool Time (in minutes) is based on 100 wafers in Infineon's database (RTT). This is an input of the simulation model and needs to be adjusted to the size of the arriving lot.

$$\mathrm{RPT} = \mathrm{scaleRPT} * \mathrm{RTT}_{lot} = \mathrm{scaleRPT} * \frac{\mathrm{RTT} * lot\ size}{100} \tag{7}$$

To give an example, we illustrate the way of one lot of product X with lot size 25 through a sequence of bottlenecks and non-bottlenecks in Villach as shown in Figure 2-5. It visits three non-bottleneck steps.

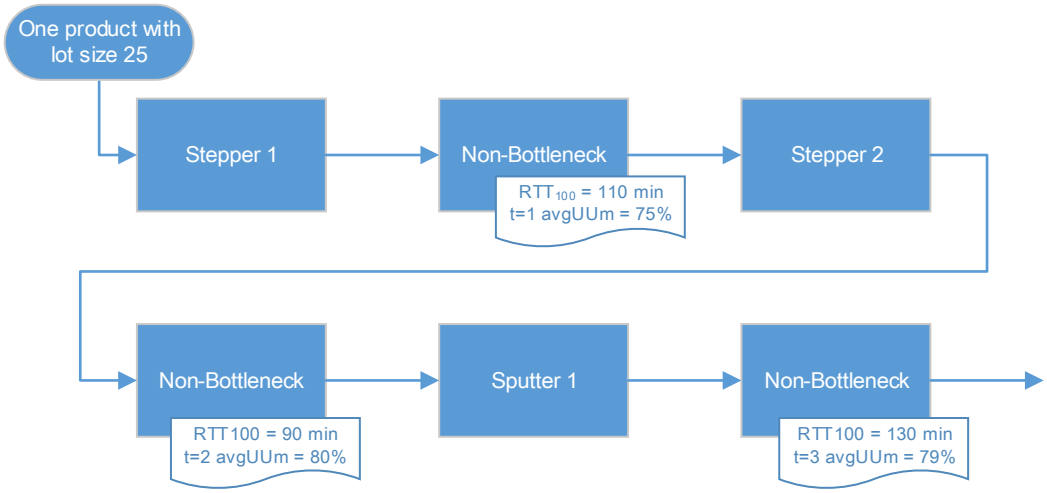


Figure 2-5 Process of product X for the delay of non-bottleneck determination example

Outside the model, from historical data, we get, for example, for week 1 a FF of 2.5 (FF_{hist}) and a utilisation of 80% (UUm_{hist}). Then, we can calculate α from this historical data, which will be used in the simulation model:

$$\alpha = \frac{(FF_{hist} - 1)(1 - UUm_{hist})}{UUm_{hist}} = \frac{(2.5 - 1)(1 - 0.80)}{0.80} = 0.38$$

Inside the model, the lot has a workroute as shown in Figure 2-5. When the lot arrives at the first non-bottleneck step in the simulation model, we have to determine the delay. First, we calculate the FF, based on the empirically determined α and the average utilisation (UUm_{avg}) observed at the steppers in the simulation model (we observe an UUm_{avg} of 75%):

$$FF = 1 + \alpha * \frac{UUm_{avg}}{1 - UUm_{avg}} = 1 + 0.38 * \frac{0.75}{1 - 0.75} = 2.13$$

Before we can calculate the non-bottleneck delay for this step we determine the raw processing time, by using the given raw tool time for that step. The factor scaleRPT is assumed to be 1.6, which is empirically estimated by the developers of the model.

$$RPT = scaleRPT * \frac{RTT_{100} * lot\ size}{100} = 1.6 * \frac{110 * 25}{100} = 44\ minutes$$

Now we can calculate the delay of the non-bottleneck step:

$$DelayNB = RPT * FF = 44 * 2.13 = 93.5\ minutes$$

The further non-bottleneck delay calculations for this example are described in Appendix E.

2.2.3 Inputs and outputs

The existing supply chain simulation model has several inputs and outputs, which we describe below. Figure 2-6 visualises the simulation model architecture. It shows that the

input file of the simulation model is connected to the databases of Infineon. After a simulation run, the output data are written into an output file.

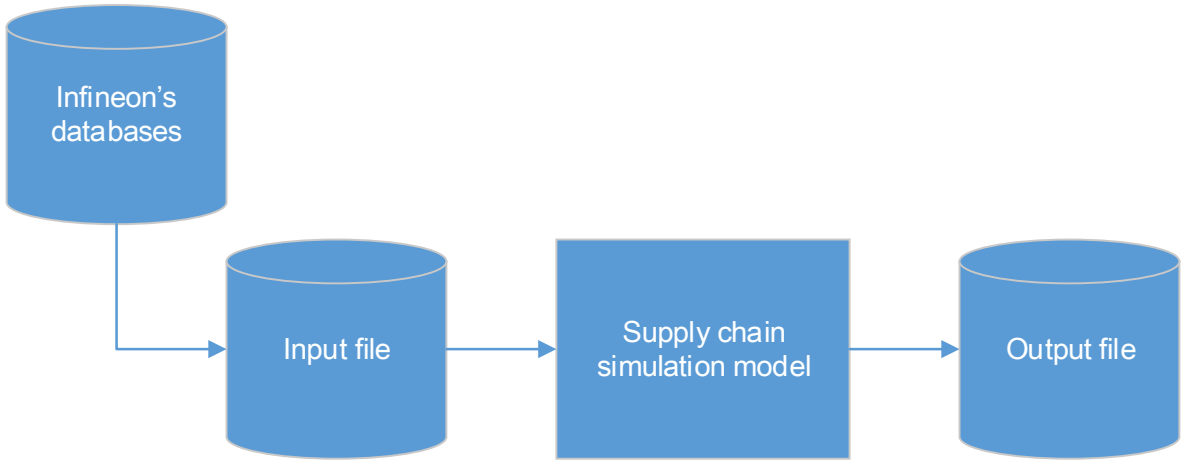


Figure 2-6 Simulation model architecture

Input

The input of the simulation model is divided in three groups of data: master data, historical data and assumed data. This input data is stored in the input file of the simulation model.

The master data consists of product master data and route master data. The product master data and its terms are explained in Appendix F. The route master data contains detailed information about the manufacturing routes. The simulation model knows which wafer needs to be processed at which facility and on which equipment by combining the work route ID of the product master data with the route master data. How this works is clarified by an example in Appendix G.

Historical data is used for parameterisation and validation of the simulation model. The input file provides the production volumes per week per basic type, the ID of the route, the week number, the lot size based on an estimated lot size per product, the number of lots, and the wafer fab facility. If there are incomplete data sets (missing data) in this input file or in the master data, such as a missing route ID for a specific basic type, they are removed from the input file for simulation and therefore not loaded into the simulation model. As a consequence, the input volume is lower in the simulation model than in reality.

Other parameters that are an input for the simulation model, more specifically for the equipment, are the uptime and the number of tools. This uptime is used to determine the downtime of the equipment. The number of tools represents the capacity of the equipment,

but as this data was not stored in the past, this number is estimated based on the current number of equipment.

Furthermore, α is used to calculate the FF of the non-bottleneck equipment as described earlier.

Moreover, the scaleRPT factor is estimated based on the sum of the Raw Tool Times from the route data for a product and the Raw Processing Time per product, available from historical data. The lot sizes are estimated based on the expected lot size per product.

Outputs

The currently used Key Performance Indicators (KPIs) are the number of wafer starts per week (WSPW), the number of layer starts per week (LSPW), the number of wafers leaving the facility per week (WOPW), the number of layers on wafers leaving the facility per week (LOPW), the utilisation of the bottleneck equipment, the WIP, the CT, and the FF. We explain these KPIs in detail in Appendix H.

Within the FE facilities in the existing model, the KPI calculation complies with the calculation stated in the technical regulations of Infineon. These regulations define how the most important KPIs are measured within each facility. It is however unclear to the developers of the existing model how lots are to be handled exactly if they visit multiple facilities within one manufacturing level (e.g. three different FE facilities within the manufacturing level wafer fab), which is important from a supply chain perspective. We indicate this shortcoming of the existing simulation model by the fact that some KPIs (the WOPW, LOPW, CT, and even WSPW as well as LSPW) in the simulation model are significantly different from the database data which is used for comparison.

2.2.4 Accuracy status

To determine the accuracy of the simulation model we need to verify the KPIs calculation in the simulation model with the KPIs calculation in the databases. As stated in the previous section (Section 2.2.3) the measurement of the WSPW, WOPW, LSPW, LOPW, and the CT is unclear. Therefore, we need to review these calculations before we can determine the accuracy status of the simulation model.

We do this by discussing the determination of the KPIs with several database experts of Infineon. We implement the discussed calculation for these KPIs in the simulation model and verify the calculations by several test runs, using the “traceln()” function of AnyLogic, and a structured walkthrough with a simulation expert. More information about how we

communicated this to the experts and how these KPIs are measured in the database is shown in Appendix I.

For validation purposes we decide to leave the WIP out, as it is proportional to the CT as described by Little's Law, and therefore it will not provide us additional insights (Brandon-Jones & Slack, 2008).

We run the existing model with a run length of 130 weeks, of which we use 52 weeks as first warm-up phase to reduce initialisation bias and 13 weeks as second warm-up phase to reduce transition bias. In the end, we have 65 weeks left for our accuracy analysis. We decide to replicate each run 10 times. How we determined these numbers and how we made these decisions is explained in more detail in Section 4.2.1 and Appendices J and K. With this approach we expect gain a good overview of the current status of the accuracy of the existing simulation model. The *accuracy* measure (Δ_{tfk}) is defined as the margin between the simulated KPIs (k) data of week t and the historical KPIs data of week t of facility f. A Δ close to 0 means a more accurate simulation model.

$$\Delta_{tfk} = \frac{(simulated\ value_{tfk} - historical\ value_{tfk})}{historical\ value_{tfk}} \quad (8)$$

This accuracy is determined for each KPI for all FE facilities, so as an overall value and for each KPI for the facilities in Villach, Regensburg, and Kulim separately. The average of Δ_{tf} over all 65 weeks over all facilities per KPI is shown in Table 2-1. Appendix L shows the average of Δ_{tf} over all 65 weeks per facility per KPI for the existing simulation model.

Table 2-1 Average overall accuracy per KPI

KPI	WSPW	WOPW	LSPW	LOPW	CT	Utilisation	FF
Δ	-2.1%	2.5%	-0.8%	4.4%	-21.9%	-3.0%	-4.8%

The difference for the WSPW, which should be equal to the historical data as this data is used as an input for the simulation model, is explained by the fact that the historical products of which data is missing, such as a route ID, are removed from the simulation input file (as described in Section 2.2.3). We analyse the ratio of missing WSPW, and can conclude that the ratio of missing values on average covers the calculated accuracy value for this KPI: an average of 2.2% is missing values versus an accuracy measure on average of -2.1%. This means, however, that we do not simulate these products, which will result in an underestimate of CT. This might be a reason for the large negative CT accuracy.

The accuracy of the CT of Kulim is the worst compared to the CT of the other facilities. This difference is clearly shown in Figure 2-7, as the two lines show a gap of about 10 days. For the other facilities, the graph looks similar. The simulated CT is for Regensburg and Kulim always lower than the actual CT.

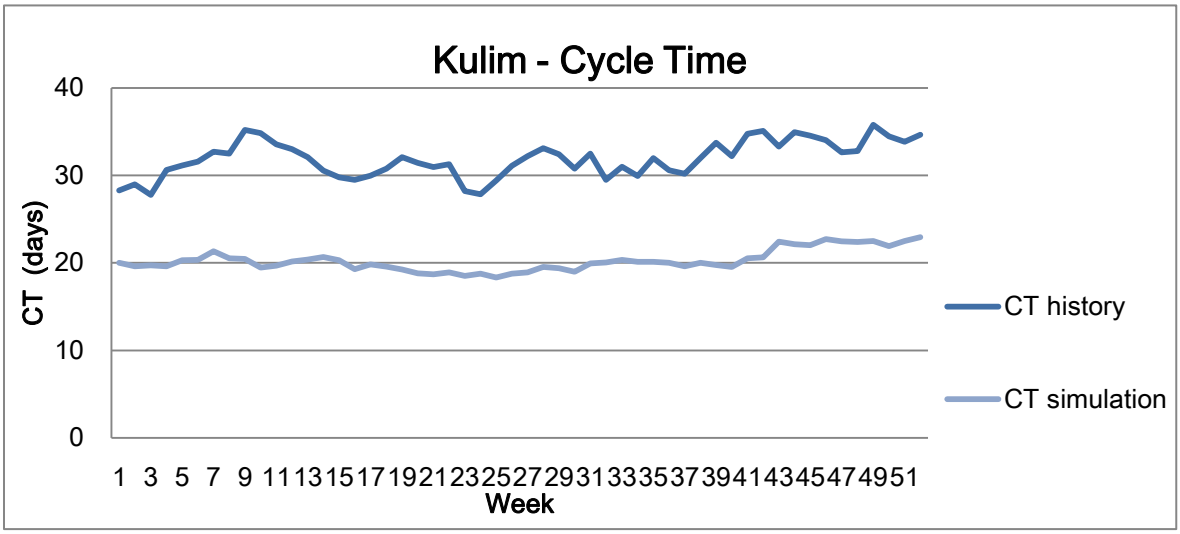


Figure 2-7 Comparison of the simulated CT and historical CT of Kulim

2.2.5 Possible problems and identified improvement areas

The existing model lacks accuracy. Discrepancies exist between the historical KPIs values and the simulated output of these KPIs. Therefore, a relationship chart is set-up to determine the relationships in the simulation model and subsequently to determine possible causes (Figure 2-8).

In Section 2.2.4 we find that the accuracy for the CT in the existing simulation model is the worst of the considered KPIs. All relations to the CT we identified in relationship chart, Figure 2-8, can be indicated as the possible cause.

Therefore, we decide to start at the left in the relationship chart (Figure 2-8): the WSPW and LSPW. As stated in Section 2.2.4 there are WSPW of certain products not considered in the simulation model due to missing data. Therefore, we conclude that we need to locate the missing data and complete them.

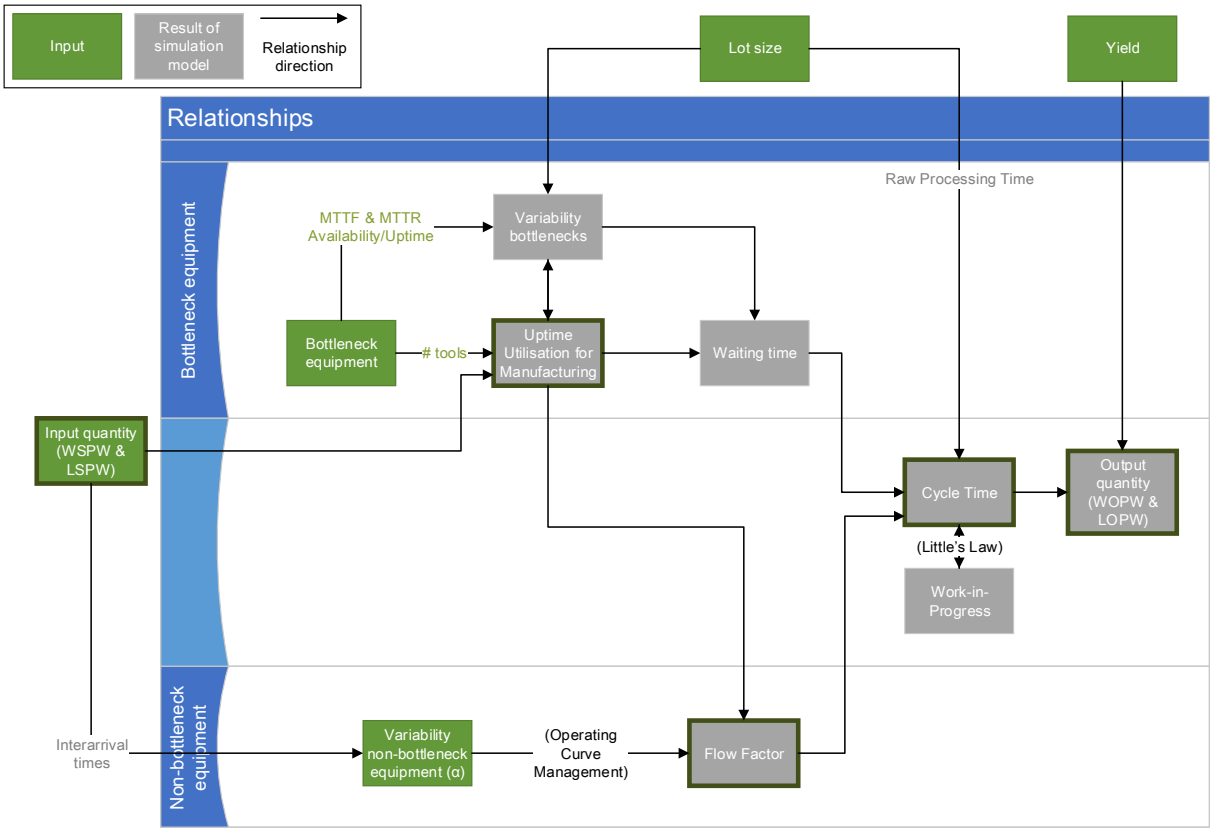


Figure 2-8 Relationship chart

The next point we come across if we move to the right in the model, are the bottleneck equipment. Currently, the bottleneck equipment is considered to be all the steppers and sputter equipment at each FE facility. It is assumed that all stepper tools at a facility can be treated equally and that the sputter tools of each facility can be treated equally. However, an analysis of the bottleneck equipment shows otherwise (Figure 2-9). In Figure 2-9 we see the utilisation weighted with the number of tools for two aggregation levels (v01: aggregation level in existing simulation model; v02: aggregation level based on dedications of machines to production processes) for the three facilities, which names are anonymised for confidentiality reasons. We observe that there are sub-clusters that show very different utilisations. We assume a machine to be a bottleneck when the utilisation is above 75%. Although this is a low threshold value to indicate a machine as a bottleneck, this value is chosen based on expert knowledge considering the criticality of the equipment; the machines are expensive and sensitive to failures.

Currently, all bottlenecks are the same for all FE facilities in the simulation model and bottlenecks cannot be easily adjusted when they change over time. We indicate this as a flexibility improvement area.

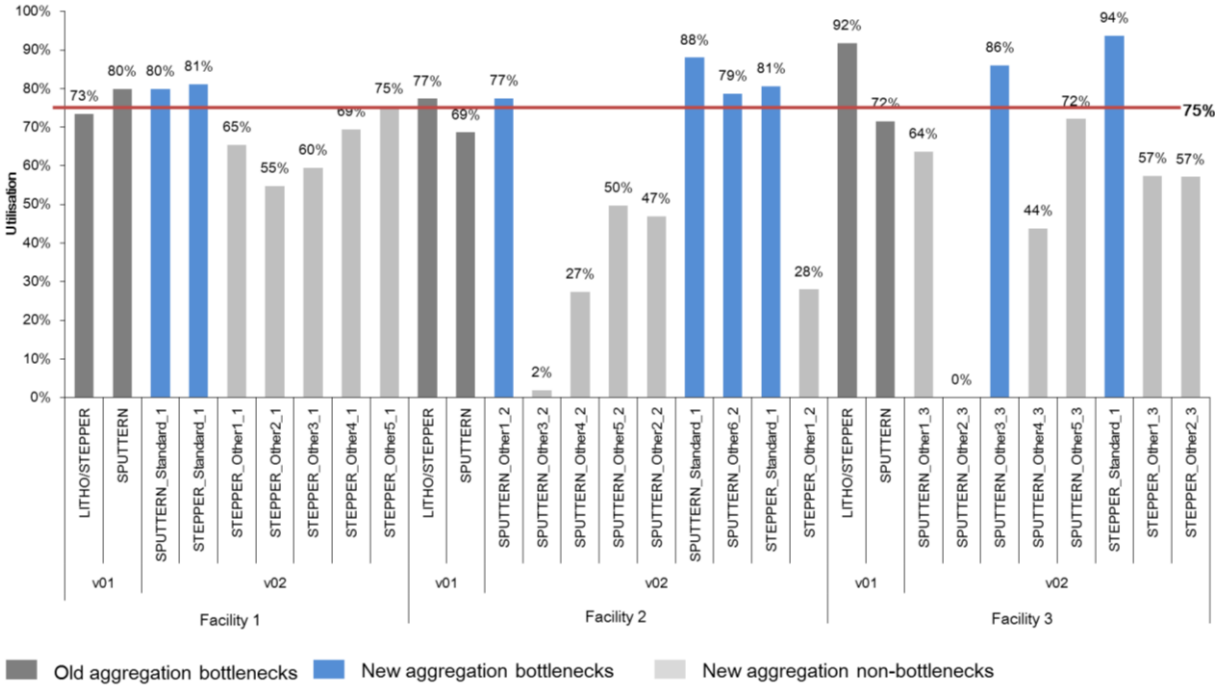


Figure 2-9 Analysis bottleneck equipment per equipment group and per location

The next point we come across if we move to the right in the relationship chart (Figure 2-8) is the variability (α). This is related to the non-bottleneck delay. We think that this can be improved by using a less aggregated variability. An α that is process group- and facility-specific can make a difference. The process group level summarises products with similar process flows.

If we move further to the right we find another possible cause. The utilisation of the sputter equipment is not considered when calculating the FF for the delay of the non-bottlenecks. This relates to the utilisation in the relationship chart. Currently, only the utilisation of the stepper equipment is taken into account. Considering the sputter utilisation as well will give a better representation of the overall facility performance.

The last point we come across if we move to the right in the relationship chart (Figure 2-8) that is related to the CT is the lot size. Therefore, we analyse the lot size mean and standard deviation over 78 weeks of historical lot size data and the simulated lot size data and conclude that the historical average lot size does not match the simulated lot size for most of the facilities (Table 2-2). Currently, this is determined based on average historical lot size per product. This has a direct impact on the CT, as the delays at the bottleneck and non-bottleneck equipment depend on the lot size. It seems to be necessary to find another suitable calculation method for the facilities in Kulim and Regensburg.

Table 2-2 Lot size mean and standard deviation analysis

		Mean	Standard deviation
Villach	Historical	39.1	0.9
	Simulated	39.3	0.8
Kulim	Historical	25.0	0.0
	Simulated	30.3	0.8
Regensburg	Historical	30.9	1.3
	Simulated	33.8	1.1

2.3 Conclusion

After an analysis of the current situation we can draw several conclusions.

We conclude that the reach represents the inventory contingency well, because stock targets are dependent on this reach (and the average demand). The Plan Load Limit represents the equipment contingency best, as the maximum planned utilisation is dependent on this Plan Load Limit (and the uptime). The reach and Plan Load Limit must be varied in the experiments simultaneously, to find the right balance between inventory and equipment contingencies.

Furthermore, we indicate improvement areas of the existing simulation model that lead to a more flexible and accurate simulation model:

- Locate and complete missing data that leads to exclusion of WSPW for simulation purposes.
- Disaggregate the selected bottlenecks into mutually exclusive subgroups and ensure the ability to adjust the facility-specific bottlenecks easily.
- Use a facility and process group specific variability as input of the simulation model to determine the non-bottleneck delay.
- Use the utilisation of all bottlenecks to determine the FF for the non-bottleneck delay.
- Reconsider the lot size calculation.

3 Theoretical framework

In this chapter, existing theory is used to build a framework for this research. Section 3.1 describes how the relevance of balancing inventory and equipment contingencies in the semiconductor industry is supported by literature. Section 3.2 focuses on supply chain simulation. In this section we describe why simulation is suitable for this research. Finally, Section 3.3 draws conclusions from this theoretical framework.

3.1 Balancing inventory and equipment contingencies

According to Hopp & Spearman (2008) three forms of buffers exist: inventory, capacity, and time. These buffers are used to correct for misaligned demand and supply losses during transformation, such as yield (Hopp & Spearman, 2008).

Bradley and Arntzen (1999) indicate that it is important to find an optimal combination of capacity and inventory what minimises costs and meets certain performance criteria. Typically, the KPIs for supply chain performance reflect the trade-off between costs and customer service (Van der Zee & Van der Vorst, 2005). The goal in the semiconductor industry is to minimise production costs and increase productivity and at the same time improving both quality and the service level. These costs are affected for a large part by yield, labour, materials, inventory, equipment and the number of wafer starts per week (Uszoy et al., 1992). Wafer fabrication facilities are very expensive. The investment needed can take up to billions of Euros (Jain, Lim, Gan, & Low, 1999; Wu, Erkoc, & Karabuk, 2005). The investment for a single machine can take up to 5 million Euros (Wu et al., 2005).

Holding (semi-)finished goods inventories is a common approach to deal with the demand fluctuations and to buffer against them in the wafer fab (Balanchandran, Li, & Radhakrishnan, 2007; Uszoy et al., 1992). However, because the customer orders and the related due dates are often not yet known when wafers need to be released into in the FE and also because the equipment of wafer fab facilities are very expensive, semiconductor companies want to keep a high throughput and equipment utilisation, while reducing CTs and inventories (Jain et al., 1999; Uszoy et al., 1992). These are conflicting objectives and therefore trade-offs have to be considered while planning.

Some models in literature developed for managing capacity (including inventory levels and excess production capacity) make use of service constraints to make sure a certain service level is maintained or reached (e.g., Gallego, Katircioglu & Ramachandran, 2006). Furthermore, most models in literature focus on minimising costs (e.g., Atamatürk &

Hochbaum, 2001; Gallego et al., 2006). Traditionally, inventory and capacity decisions are made independent from each other, although this can better be done simultaneously as it leads to a better financial outcome (Atamatürk & Hochbaum, 2001; Bradley & Arntzen, 1999). Analytical models that combine these decisions, and other capacity decisions, are mostly based on Newsvendor-Style models (Wu et al., 2005). For instance, Van Mieghem and Rudi (2002), developed a Newsvendor Network model in which multiple products, echelon inventory and multiple processing points can be considered. We found no specific literature on integrating inventory and capacity decisions by the use of a simulation model.

By planning inventory and capacity decisions simultaneously, it is possible to balance the conflicting objectives mentioned. As stated before, this balance should be obtained by minimising costs and maintaining a certain service level.

3.2 Supply chain simulation

Spreadsheet and queueing models are used when basic questions about a system need to be answered. However, in complex systems, such as in a supply chain in the semiconductor industry, discrete-event simulation is needed to answer more detailed questions (Chance, Robinson, & Fowler, 1996). Supply chain simulation can be used as a tool for decision-support, as well as for experimenting with different scenarios. A lot of different scenarios are thinkable in complex supply chains and therefore one should pay attention to the simplicity and transparency (Van der Zee & Van der Vorst, 2005). The simulation model itself should be easily accessible, it should visualise the processes, by for instance using graphs, and give insight in the key decision variables. In this way decision makers can gain insight and oversight with regard to different supply chain scenarios and gain trust in the simulation model. This trust is needed to make the parties involved more likely to accept the outcomes of the simulation study (Chance et al., 1996; Van der Zee & Van der Vorst, 2005). Simulation provides further benefits. By the means of simulation, the real system does not have to be interrupted for experiments (Terzi & Cavalieri, 2004). Moreover, the costs of experimenting with different scenarios are negligible, whereas experimenting in the real world this can be pretty costly and risky (Fowler & Rose, 2004; Terzi & Cavalieri, 2004; Van der Zee & Van der Vorst, 2005; Yuan & Ponsignon, 2014).

In literature, the most mentioned simulation method for supply chain simulation is discrete-event simulation (e.g. Duarte, Fowler, Knutson, Gel & Shunk, 2000; Fowler & Rose, 2004; Morrice & Valdez, 2005). With this method many manufacturing and planning activities can be captured (Yuan & Ponsignon, 2014), which is important for this research. Discrete-event

models are dynamic, stochastic, and discrete. Dynamic, since the simulation model represents a system as it evolves over time. Stochastic, because probabilities are used for certain inputs, for instance for machine uptime or customer orders. Discrete, as the system changes in a countable number of points in time (Law, 2007). By discrete event simulation, lots can be tracked through all facilities in the supply chain (Fowler & Rose, 2004).

Before starting the modelling we need to ask ourselves two major questions: do we want to reuse the simulation model? And at what level of detail do we want to model? Section 3.2.1 and Section 3.2.2 explore these decisions respectively. Afterwards, the focus is drawn on wafer fab simulation (Section 3.2.3). Here the existing literature is reviewed and linked to this research. This is followed by a description of several validation and verification techniques for verifying and assessing the validity of simulation models (Section 3.2.4).

3.2.1 Reusability

If the purpose of a simulation model is to analyse different scenarios a reusable simulation model might be preferred. This reusability can be achieved in different ways: reusing small parts of codes, reusing larger components or reusing complete models (Robinson, Nance, Paul, Pidd, & Taylor, 2004). A reusable simulation model does not only enable to model more rapidly after the first model is built, but also reduces costs and modelling errors (Van der Zee & Van der Vorst, 2005; Yuan & Ponsignon, 2014).

This is also the target for the supply chain model of Infineon. It is however hard to make a reusable simulation model on a high level (Robinson et al., 2004). To create a reusable simulation model on a supply chain level, flexibility should be incorporated to be able to change the simulation model easily in case of changes in the manufacturing environment, or when other scenarios need to be studied, in this case for example another contingency type. This flexibility also reduces the maintenance needed for the simulation model (Chance et al. 1996). Moreover, Law (2007) states that object-oriented simulation supports reusability, as the developed objects can be reused or adjusted easily.

However, there are also obstacles for the development and usage of a reusable model. Modellers might not want to develop such a simulation model, because it costs more time and others benefit from it. Furthermore, it can be hard to rely on someone else's code and if the code is trusted, it might take time to fully understand the code someone else wrote (Robinson et al., 2004). The result of these obstacles could contradict the advantages of reusability; it could lead to more need of time and costs. Nonetheless, the last obstacle can probably be reduced by using good documentation.

Although aiming for reusability entails several obstacles, Infineon is confident they can reduce or overcome them for their supply chain simulation model. Infineon already has a lot of experience with complex reusable models on the work centre level and the facility level (see Section 2.2.1 for more information on the levels).

3.2.2 Level of detail

Finding the appropriate level of detail is important in supply chain simulation modelling (Fowler & Rose, 2004). For a less detailed simulation model also terms as reduced models or simplified models are used in literature. The term 'simplified simulation model' is defined in Section 2.2.2. A reduced simulation model or a model with a lowered level of detail is for this research defined the same. According to Sprenger & Rose (2011, p.453) a model can be simplified by, among others, "removing unimportant components of the model, using random variables to replace parts of the simulation model, considering less detail for the range of variables of the simulation model, and combining components of the simulation model into new and simpler components".

Simplification of the simulation model has several advantages and disadvantages. A simplified simulation model is often more easily understood than a complex simulation model. Furthermore, a simplified simulation model runs faster, so results can be obtained more quickly. However, the pitfall with simplified models is that if the simulation model is made too simple, important factors of the system might be omitted. This can lead to inaccurate results, which affects the comprehensibility and acceptance of the model. Moreover, by simplifying the model too much, the experiments that can be conducted with it are being limited (Brooks & Tobias, 2000).

It is important to define the goal and use of the simulation model before deciding on the level of detail. For two cases Brooks & Tobias (2000) compared a complex and a simplified simulation model. They conclude that a simplified simulation model, in which equipment and buffers are both aggregated, whenever possible, performs better with respect to the running time and the comprehensibility of the simulation model. In addition, in the complex simulation models more errors in the construction are found (Brooks & Tobias, 2000).

These results differ from the outcome of the research of Jain et al. (1999). They investigated what the level of detail for supply chain simulation modelling in the semiconductor industry should be by testing two levels of detail: supply chain simulation with only bottleneck equipment modelled in detail, and supply chain simulation with all equipment modelled in detail. In the bottleneck-only simulation model, the non-bottlenecks are modelled as delays,

considering the processing times, set-up times, waiting times and travel times. The bottlenecks are determined for each route based on their utilisation. In the detailed model all steps are modelled, where the equipment and operators are considered as constraints. Their conclusion is that a detailed simulation model gives more accurate results than the bottleneck-only simulation model when looking at typical KPIs for semiconductor supply chains (Jain, Lim, Gan, & Low, 1999). However, the results are not validated against a real system, because a fictional case is used. The results of the detailed simulation model are used as a reference level and the results of the bottleneck-only simulation model do not comply.

Jain et al. (1999) modelled a process at a semiconductor manufacturer and Brooks & Tobias (2000) have researched cases outside the semiconductor industry. Both have different outcomes and we think that the results of the research of Jain et al. (1999) need to be questioned although the setting of their research is more similar to that of this research. The supply chain simulation model built by Infineon for contingency optimisation has a specific purpose and because of that very detailed, complex modelling is not necessary. Furthermore, the running time of the simulation model should not be too long as it would be useless for rapid scenario analysis. Moreover, the results should be communicated to other departments. A high level of detail might increase the credibility of the simulation model, but it takes more time consuming building, running and maintaining the simulation model (Brooks & Tobias, 2000; Fowler & Rose, 2004; Sprenger & Rose, 2011). Therefore, it is important to make the simulation model comprehensible and accessible for scenario analysis by building a more simplified simulation model of reality and to pay extra attention to the accuracy of the simulation model to make sure it is valid.

3.2.3 Wafer fabrication simulation

Simulation is an important tool for analysing and modelling wafer fabs and is described in different papers (e.g., Chien et al., 2011; Hung & Leachman, 1999; Peikert, Thoma & Brown, 1998; Piplani & Puah, 2004; Jain et al., 1999; Rose, 2004; Rose, 2007; Sprenger & Rose, 2011; Uszoy et al., 1992). The modelling of wafer fabs is difficult, because of the complexity described in Section 1.2. Sprenger & Rose (2011) reviewed existing approaches and build their research upon them. Therefore we decided to describe their approach in more detail. Moreover, the research of Peikert et al. (1998) is described in more detail, because their approach differs from the approach of Sprenger & Rose (2011) and it is close to the approach used by Infineon currently.

Sprenger & Rose (2011) have reduced the complexity of a wafer fab simulation model by modelling as few as possible pieces of equipment in detail and replacing the other equipment by delays. Equipment with a high utilisation is not replaced by a delay, because it has a large impact on the wafer fab and the behaviour of the lots. Furthermore, re-entries at bottleneck machines are considered as well, before entering the queue at the bottleneck equipment again (Figure 3-1).

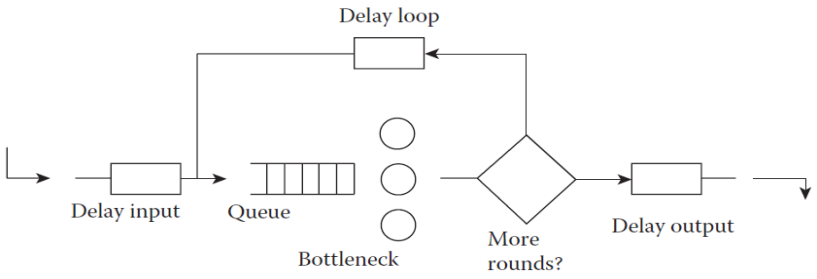


Figure 3-1 Simple simulation model (Sprenger & Rose, 2011, p.454)

The authors have used the number of machines, a MTTR and a MTTF, processing times, set-up times and set-up rules as parameters for the bottleneck equipment group. The focus of their research is on two approaches to determine the delay in the loop and comparing them to the usage of static distributions for these delays, where all lots of a product are delayed using the same distribution. The two approaches are a delay approach and an interarrival time approach. In the delay approach, a delay distribution is determined for each product depending on the number of lots that need to be processed by the non-bottleneck equipment at any particular moment. In the interarrival time approach the delay is dependent on the interarrival distribution at the bottleneck queue of the lots from the loop, also depending on the number of lots that need to be processed by the non-bottleneck equipment at that moment. For this interarrival distribution three options are explored: based on the interarrival time between all lots, interarrival times between lots with all combinations of products, and interarrival times between lots of the same product. The first option has proven to be suitable in most cases.

These two approaches have been tested by comparing the operating curve of a complex simulation model to the operating curve of the simple simulation model. Both approaches have been proved to work better than using a static delay distribution. The delay approach seems more applicable in case of changed product mix than the interarrival time approach. However, the interarrival time approach needs more research and improvement. Therefore, we conclude from their research that a dynamic delay distribution leads to a better

performance of the simplified simulation model than a static delay distribution. (Sprenger & Rose, 2011)

A short coming of the study of Sprenger and Rose (2011) is that only six products are considered and only one single facility is modelled. This leads automatically to a simpler simulation model compared to the simulation model used for supply chain simulation at Infineon and could lead to different results when using the approach used by Sprenger and Rose (2011). Furthermore, they measured the performance of their simplified simulation model by using a complex and very detailed simulation model as a reference instead of using real data, which Jain et al. (1999) also did in their research. This was not possible, because they used a fictional case. Furthermore, the detailed simulation model was not validated against real data. However, a complex simulation model cannot always be assumed to be generating the right output values, especially when it is not validated against real data, because errors can be made in the construction (Brooks & Tobias, 2000).

Peikert, Thoma & Brown (1998) have determined the non-bottleneck delay differently and used a real case; the wafer fab in Dresden of Siemens Microelectronics Center (is now Infineon). They modelled a single wafer fab and focussed on modelling only bottlenecks as well. Their simulation model looks similar to the simulation model of Sprenger & Rose (2011) captured in Figure 14. The lithography equipment, the steppers, are considered as bottlenecks and are therefore modelled in more detail; the rest of the equipment is captured in a ‘black box’ as dummy equipment with an unlimited number of servers. The delays have been determined by first calculating the processing time by adding all Raw Processing Times of the actual production steps. The delay time is subsequently determined by multiplying the Raw Processing Times of each step by a lead time factor, that is derived from historical data and correspond to the real FFs (Peikert et al., 1998). For the delay in the loop, Peikert et al. (1998) have used lead time factors for different product groups and the processing time has been drawn from a triangular distribution. These product groups have been determined by aggregating products with similar process routes at a combined wafer start rate. Rework has only been modelled for the stepper equipment and is based on the actual rework rate. Scrap is also considered for the stepper equipment by an average scrap rate based on historical data and is incorporated in each operation at this equipment. For the ‘black box’ equipment this was incorporated for all operations (Peikert et al., 1998).

This is almost the same way the delays are calculated in the current simulation model of Infineon. The only difference is that for the simulation model of Peikert et al. (1998), only the lithography equipment is modelled as bottlenecks. However, there might be other or more

bottlenecks. Therefore, these are modelled in the current simulation model of Infineon as well. The loop used in the simulation model of Peikert et al. (1998) and Sprenger & Rose (2011) is also applicable for the current simulation model of Infineon. Although the bottleneck and non-bottleneck equipment is modelled on different layers, a possible re-entry is defined by the routes in the master data. Product aggregation is also used in the current simulation model of Infineon, for which the routes on this level are still correct. Scrap is already incorporated in the simulation model of Infineon as yield. Rework is however not considered in the current simulation model, because the rework time is negligible at the currently modelled bottleneck equipment.

The simulation model of Peikert et al. (1998) is validated against historical data and obtained good results. The stepper utilisation has an accuracy of 90% and the CT an accuracy of 97% (Peikert et al., 1998), however, the exact calculation of this accuracy is unclear. Nevertheless, we conclude that way of delay modelling for the non-bottlenecks of Peikert et al. (1998) has proven to be working and that their method is more comparable to the current simulation model of Infineon than the approaches used by Sprenger & Rose (2011). Therefore, we decide to stick to this approach.

3.2.4 Verification and validation

There are different ways to determine whether a simulation model is 'good enough' described in literature. Sargent (2013) describes seventeen techniques for the validation and verification of a simulation model. Generally, a combination of these techniques is used. After a literature review on verification and validation (e.g., Kleijnen, 1995; Law, 2007; Sargent, 2013) and a discussion with simulation experts of Infineon to check which techniques fit this project best, we decide to use five of the described techniques by Sargent (2013) to validate and verify the further developed supply chain simulation model: face validity, historical data validation, internal validity, operational graphics, and structured walkthrough. The use of these techniques is feasible with the resources that we have to our disposal and we believe these five will give a sound foundation for the verification and validation of our model. We explain each of the five techniques below:

- *Face validity* is the extent to which individuals that are knowledgeable about the system think the simulation model and its behaviour represent reality (Sargent, 2013). Using animation is a good example to assess face validity (Kleijnen, 1995).
- *Historical data validation* is the extent to which historical data of the system match the simulated data, like a comparison of KPIs. For this end, some historical data must be

used as inputs of the simulation model for testing (Sargent, 2013). This is also called 'trace driven simulation' (Kleijnen, 1995). The required accuracy depends on the purpose the simulation model is intended for. Numerical statistics, such as the sample mean, sample variance and sample correlation, can be used for this comparison, as well as statistical tests. Furthermore, the comparison could be done by graphical representations, like box plots, histograms and spider-web plots (Law, 2007).

- *Internal validity* is the extent to which the simulated output varies when several replications of a stochastic model are made (Sargent, 2013). Before starting the experiments it is necessary to determine the number of replications needed to make sure that the outcomes lie within a sufficiently low confidence interval, based on a relative error (Law, 2007).
- *Operational graphics* are a graphical representation of different performance measures during the simulation run, so that the development can be checked over time. This is done to check if the simulation model and the KPIs behave correctly. (Sargent, 2013)
- In a *structured walkthrough* the model developer shows the simulation model and its layers and objects to a peer group to check if the logic of the model is correct and if the necessary accuracy is obtained (Sargent, 2013). Furthermore, the simulation model's assumptions need to be assessed on correctness and completeness (Law, 2007).

In literature several KPIs that are typical for the semiconductor manufacturing systems are mentioned, such as the utilisation, WIP, FF, CT, and service level (Chien et al., 2011; Duarte et al., 2002; Rose, 2007; Yuan & Ponsignon, 2014). We need to add the WSPW, WOPW, LSPW and LOPW to these KPIs to make sure the wafers enter and leave the facilities correctly in the simulation model as we model a network of facilities.

3.3 Conclusion

Although literature mentions that considering capacity and inventory levels simultaneously is financially beneficial and simulation is preferred over analytical models in case of complex systems, not much research has been done to balance these buffers by using simulation. We conclude, based on the advantages and disadvantages of simulation described in Section 3.2, that this is the right method for this research on the supply chain level and that discrete-event simulation fits the purpose of the project best.

It is important to consider if the simulation model needs to be reused and on what level of detail should be modelled. Reusability is important for Infineon, because the simulation model is needed for analysing various scenarios. By the use of object-oriented simulation and by introducing flexibility this can be reached. Furthermore, we conclude that it is not necessary to model the supply chain on a detailed level to reach good results. A bottleneck only approach works well, which means that only the bottleneck equipment is modelled in detail.

We have found five verification and validation techniques that are important for this project: face validation, historical data validation, internal validation, operational graphics, and structured walkthroughs. The most used KPIs in this industry, which we choose to verify and validate the simulation model with, are the utilisation, FF, CT, service level, WSPW, WOPW, LSPW, and the LOPW.

4 Updated simulation model

In this chapter we develop the simulation model that provides us the flexibility and accuracy which will help us to answer the main research question (Section 1.6). First, we describe the conceptual design and implementation of the updates in the existing supply chain simulation model (Section 4.1). Section 4.2 follows with the set-up and results of the validation experiment. We draw conclusions in Section 4.3.

4.1 Concepts and implementation

In Chapter 2 we discussed several improvement areas to the existing supply chain simulation model. This section outlines the conceptual design and implementation of the indicated improvement areas of the existing supply chain simulation model in Chapter 2 and is divided into 4 sub-sections. Section 4.1.1 describes the adaptions made to the wafer fabrication object introduced in Section 2.2, to improve its accuracy and flexibility. Section 4.1.2 explains how the rest of the supply chain is represented for the contingency experiments. Section 4.1.3 lists the assumptions made and Section 4.1.4 describes the verification of the simulation model.

4.1.1 Accuracy and flexibility improvement areas of the wafer fab object

In this sub-section we describe the required changes and additions to the existing wafer fab object of the simulation model, as indicated in Section 2.2.5. This section focuses on the changes that are necessary for the validation experiment.

Facility-specific bottlenecks

In Chapter 2 we concluded that the bottlenecks in the existing simulation model were too much aggregated. We decide to start with this improvement, because it focuses on improving both the accuracy and on the flexibility of the simulation model. These two aspects are both stated in the objective of this research (Section 1.4). The inclusion of facility-specific bottlenecks brings more detail into the simulation model, which we expect to lead to a more representative estimate of the fab utilisation. Therewith, we expect the CT to come closer to its historical value as well, because they are linked both directly (waiting time depends on utilisation) and indirectly (utilisation of the bottlenecks is used to calculate the delay at the non-bottlenecks) (see relationships in Figure 2-8). For this matter, we integrate a bottleneck object in the updated simulation model that is adjusted automatically when the bottlenecks change in the database.

To create the flexibility of changing the bottlenecks we have chosen to set the bottlenecks as an input and read them into the simulation model on initialisation. We link the input file to the corresponding database, which enables us to update the bottlenecks anytime and without much effort. Moreover, we can adjust the routes of the wafers accordingly. We ensure that the parameters of these bottlenecks (number of tools and availability) are updated in the model every simulated week. The implementation in the simulation model is described in Appendix I. Furthermore, we adjust the insertion of R&D (non-sales production) lots to these new bottlenecks in such a way that the share of R&D lots can be controlled individually for each bottleneck.

Delay determination non-bottlenecks

According to the relations we visualised in Figure 2-8, the CT depends on the utilisation of the bottlenecks and variability (α) of the production (the quality of the non-uniformity of the production, see Section 2.1.2). These factors contribute to the delay determination at non-bottleneck steps, which is based on the Operating Curve theory (see Section 2.1.2 and Appendix C for more details). We consider the following two improvement areas we identified in Chapter 2: the estimation of α and the utilisation used to calculate the non-bottleneck delay. By bringing more detail to the existing simulation model on these areas we expect that the FF will be estimated better as well, which leads a better estimate of the non-bottleneck delay and therefore to a more accurate CT (see relationships in Figure 2-8).

To estimate α more accurately, we choose to use process group specific values of α. Alternative aggregation levels are facility specific values, as used in the existing simulation model, or on material number level (the product number used within Infineon), which would be very detailed. The process group level summarises products with similar process flows. It is therefore more detailed than the facility level, but requires less data than the material number level. In this way, we ensure that the simulation model does not have to store too much data which would affect the performance of the model negatively and we are still able to bring more detail to the model to provide a good representation of reality. We include these new values in the input file and adjust the existing simulation model in such way that it uses this new input to determine the non-bottleneck delay.

To achieve this, we extract the FF per week for each process group from historical data and calculate the historical average utilisation of each bottleneck per facility per week. We do this by creating a query in the input file. α is then determined as we described in Section 2.2.2. Unfortunately, not for all process groups an α value is available every week. We have

decided to solve this problem by taking the historical average α per week per facility in case of a missing α per process group. The implementation is described in more detail in Appendix I.

Besides improving the estimation of variability parameter α in the simulation model, we want to make sure that the utilisation of all bottlenecks is taken into account when determining the FF for the delay calculation of the non-bottlenecks. The rationale behind this is that including more bottleneck stations gives a better representation of the overall fab.

To ensure that the utilisations of all bottleneck equipment are taken into account when determining the non-bottleneck delay, we store the utilisation of each bottleneck in a separate statistics dashboard. When the delay needs to be determined (i.e. when a lot arrives at a non-bottleneck step), the simulation model takes the average utilisation by summing the utilisation of all bottlenecks in the statistics dashboard and dividing this number by the number of bottlenecks of that facility. We describe this in more detail in Appendix I.

Facility-specific lot size determination

As shown in Figure 2-8, the lot size both indirectly and directly influences the CT. As shown in Table 2-2 the lot size determination method, which is the same for all facilities in the existing model, does not represent reality for all facilities (Kulim and Regensburg). Therefore, we search a suitable calculation method for determining the lot sizes for each facility separately.

We come up with facility-specific lot size determination methods by trying logical rules and by setting up rules based on information from experts working at the FE facilities. We discover that in Kulim a standard lot size of 25 is used. If the number of wafer starts cannot be divided by 25, the remainder is considered as one lot. For Regensburg we encounter that the lot sizes are a mix of lot sizes of 25 and 50. We analyse the historical lot sizes over 78 weeks, from fiscal week 401 to fiscal week 525, and we conclude that a mix of 2/3 of the products with lot size of 50 and 1/3 of the products (aggregated on material number level: product number used within Infineon) with lot size of 25 represents reality best. We compare the result of the new lot size determination to the historical lot size over 78 weeks of Regensburg in Figure 4-1. The graph shows that the new method reproduces reality to a high degree.

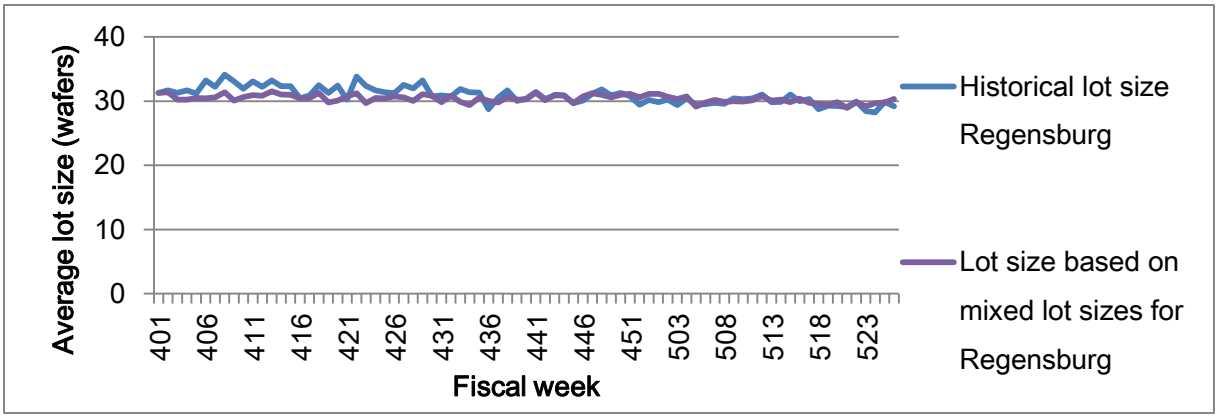


Figure 4-1 Mixed lot sizes analysis for Regensburg

We calculate the lot size with the updated methods for Kulim and Regensburg and calculate the mean and standard deviation of the lot sizes over 78 weeks (Table 4-1). The mean and standard deviation are closer to reality than when using the previously used method in the existing simulation model (Table 2-2).

Table 4-1 lot size mean and standard deviation of the existing and the updated method and the historical lot size

		Mean	Standard deviation
Kulim	Historical lot size	25.0	0.0
	Lot size updated method	25.0	0.0
	Lot size existing method	30.3	0.8
Regensburg	Historical lot size	30.9	1.3
	Lot size updated method	30.4	0.6
	Lot size existing method	33.8	1.1

We describe the implementation in the input file of the simulation model in Appendix I.

Other adjustments

We use data that was collected for other purposes than simulation. Therefore, we have to validate the data and make adjustments to fit the data to our own purposes. For example, the raw tool times in the database are based on 100 wafers. However, in our simulation model, not always 100 wafers are produced at once. This leads to deviations in the simulated raw tool time and the raw tool time in the database when we recalculate the raw tool time for each lot (see Section 2.2.2, Equation 7), e.g. because of different set-up time assumptions. We introduce factors, for example a raw tool time factor, to make up for these discrepancies. We explain other cases below and note that using factors is only a temporary solution to be used during the time data tailored to simulation purposes is not available.

We study the raw processing time per facility per week in the simulation model and conclude that these values do not match the historical raw processing time per facility for those weeks in Kulim and Regensburg. Since transportation time between equipment differs per facility, we decide to make the parameter scaleRPT, which translates the raw tool time of a product at the equipment to a raw processing time (more information in Section 2.2.2), facility-specific.

Most sputter equipment consists of cluster tools, which consist of chambers. These chambers can be used in parallel for production. Since (simulation) modelling of cluster tools is complex and we have limited time, we decide to use another approach: we include these chambers in the number of tools of that bottleneck by factors, so increase the capacity. For example, if the number of chambers is three, we use a capacity factor of 3 to increase the number of tools of the equipment. Instead of simulating 1 tool with 3 chambers we simulate 3 tools with 1 chamber.

4.1.2 Representation of sort, assembly, and test

To be able to simulate the material flow from the release into the fab until the arrival at the DCs for our contingency experiments, we use a CT approximation approach for the manufacturing levels (see Section 1.1.2) sort, assembly, and test. This approximation is based on the planned CT and the historical deviation from this planned CT, the so called *CT spread*.

This is an empirical simplified solution to include CT variability at these manufacturing levels and is chosen over a theoretical CT distribution. We made this decision because determining a theoretical distribution for all products on material number level (the product number used within Infineon) or another less aggregated level leads to many different distributions with different parameters, which is a lot of work to implement in the simulation software. Furthermore, we are of the opinion that one theoretical distribution for all products does not represent reality to such a degree that we are able to gain representative results when we use this method to represent the CT for sort, assembly, and test in our experiments. We think that our empirical approach is therefore the best fit for our purposes.

We aggregate the CT spread for sort on process group level and the CT spread for assembly and test on package level. The process group level summarises products with similar process flows. The package level summarises chips that are assembled the same way. These are the aggregation levels Infineon typically works with in the FE and BE (as explained in Section 1.1.2). We use the following approach:

1. A deviation from the planned CT is observed over a period of 13 weeks (the third quarter of Infineon's fiscal year 2014/2015) as for example shown in Figure 4-2. We obtain this data from Infineon's reporting system.

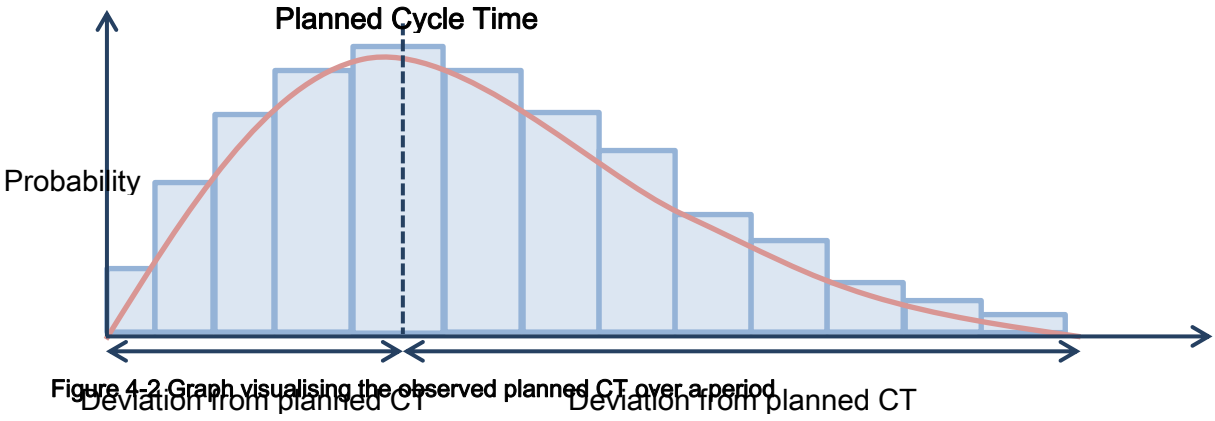


Figure 4-2 Graph visualising the observed planned CT over a period

2. We extract quantiles 0, 5, 25, 50, 75, 95, and 100 from the database, which makes it possible for us to draw the CT spread as a probability density function.
3. We translate the probability density function into a cumulative distribution function as, for example, shown in Figure 4-3. If, for instance, the CT spread value at Q5 is -2, we can conclude that 5% of the considered lots finished the corresponding manufacturing level two days faster than planned.

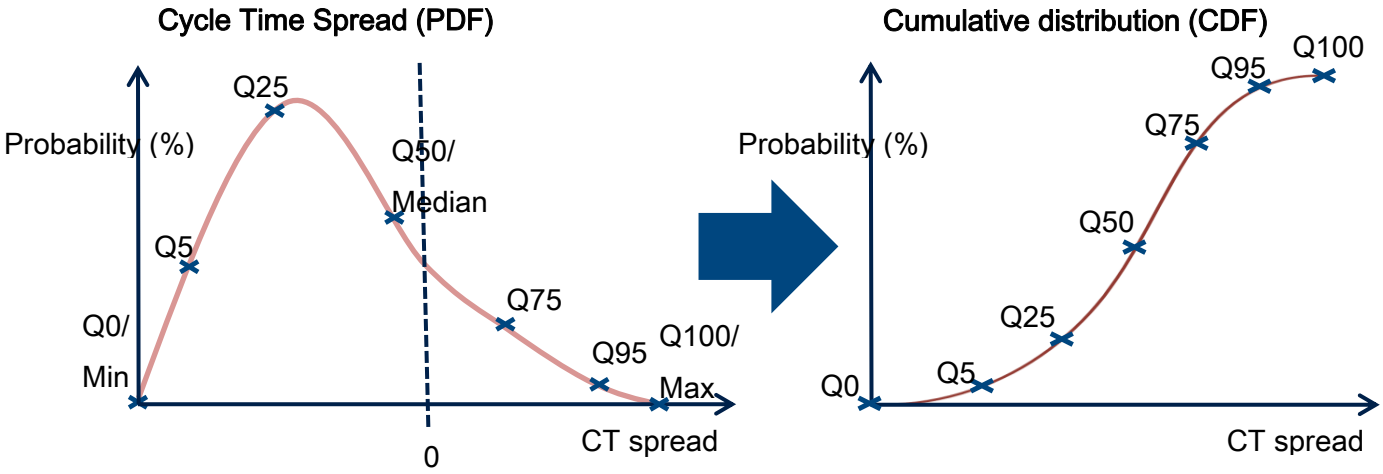


Figure 4-3 Cycle Time Spread PDF and CDF example

4. When a lot arrives in the simulation model, AnyLogic draws a random number between 0 and 100 by using the uniform distribution to determine which quantile we consider. For example, if the simulation model draws the number 10, we want the value of quantile 10. Therefore, we consider quantile 5 and quantile 25, because 5 and 25 form the limits of the interval quantile 10 belongs to.

5. We interpolate between the quantiles associated with that product to get an estimation of the spread. If, for example, the random number drawn is 10 (uniform), we use the values of Q5 ($Qmin_{value}$) and Q25 ($Qref_{value}$) to interpolate by using the formula:

$$CT\ spread\ = \frac{Qref_{value}\ -\ Qmin_{value}}{Qref\ -\ Qmin}\ *(uniform\ -Qmin) + Qmin_{value} \quad (9)$$

   If the value of Q5 is -2 and the value of Q25 -0.5 then the CT spread used for the CT approximation is -1.625.
6. We add the CT spread to the planned CT and represent the obtained time by a delay in the simulation model.

4.1.3 Assumptions and simplifications

Further assumptions we made, besides those mentioned in Sections 4.1.1 and 4.1.2, when developing the simulation model are:

- Transportation time and disturbances between the wafer fab locations can be neglected.
- The factories operate 24 hours/day and 7 days per week. Downtimes are considered in the availability of the bottlenecks.
- There are always enough operators available at the machines. Operators are not modelled.
- The wafer lot size determined at the first facility remains constant within the FE.
- A share of 5% utilisation at each modelled bottleneck due to R&D lots (non-sales production) is considered. This is a realistic value based on expert opinion.

4.1.4 Verification

As described in Section 3.2.4 there are multiple ways to verify the simulation model. The technique we use for verification is a structured walkthrough. We discuss the simulation model with peers during biweekly supply chain simulation meetings to assess the assumptions and to check the completeness and logic of the simulation model. Moreover, we use the “traceln()” function of AnyLogic to check the parameterisation and calculations during the simulation run. We solved all the errors in the simulation model that the model showed us in the console and the simulation model runs smoothly.

4.2 Validation experiment

To determine the accuracy of the improved wafer fab part of the simulation model by the historical data validation technique, we perform a validation experiment (Section 4.2.3).

Before we start this experiment we use other, less time consuming, quantitative and qualitative, validation techniques first. We use operational graphics, such as line charts and histograms in which we represent both the simulated and the historical (empirical) KPI values (as shown for the WOPW in Regensburg in Figure 4-4).

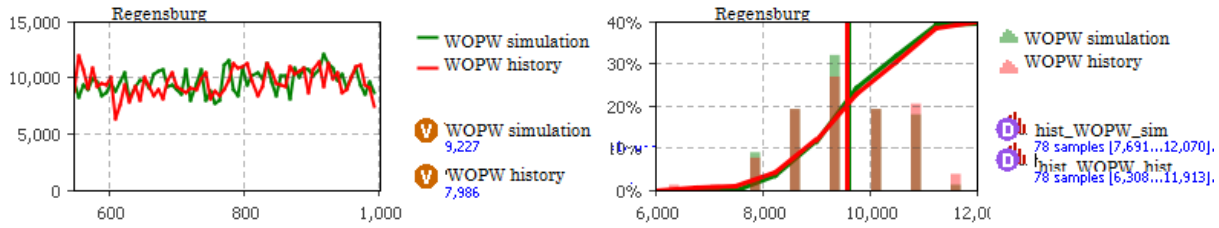


Figure 4-4 Example of operational graphics in the simulation model

Furthermore, we check the model with the face validity technique by using animation when we run the model. By this end, we can see the lots move through the different facilities and their corresponding bottlenecks.

Section 4.2.1 discusses how to set the warm-up period, run length, and the number of replications used for the validation experiment. Section 4.2.2 describes the input data used for the validation experiment and Section 4.2.3 outlines the obtained results and therewith, the accuracy of the simulation model.

4.2.1 Warm-up period, run length, and number of replications

To reduce the initialisation bias, it is important to define a warm-up period. Furthermore, it is important to define the number of replications, to gain statistically significant results. We do this for all facilities for both the existing and the updated simulation model, to be able to quantify the improvement of accuracy of the updated simulation model due to the adjustments described in Section 4.1.1 as well.

The idea behind using a warm-up period is to run a simulation model until it reaches a realistic initial state. The data obtained in this period is deleted for analysis.

Hoad, Robinson, and Davies (2008) did an extensive literature review on methods to determine the length of the warm-up period. One of the 42 methods they found is a graphical method, which is also extensively described in Law (2007): Welch's graphical procedure. This method is a well-known and widely used method and according to Law (2007) the

simplest and most general technique. Another way reduce the initialisation bias is to run the model for a longer time (Hoad et al., 2008).

Generally speaking, we should use historical data to determine the warm-up period with Welch's graphical procedure, but we have not enough historical data available. Furthermore, Infineon grew a lot during the last years and faced a lot of fast ramp-ups. Therefore, we cannot use the data of the 78 weeks we have available to determine the warm-up period; it will not be possible to reach a steady state. However, we need a warm-up period to fill the system and to cope with an initialisation bias. Therefore, we decide to use two warm-up phases for our warm-up period.

We want to use a first warm-up phase, because we start our simulation with an empty system and want to fill the model. For this, we use the well-known method of Welch. As there is no steady state behaviour in our simulation model when we use historical data, we can run the model for a longer time to make the initialisation bias effect negligible (Hoad et al., 2008). We use this approach for our second warm-up phase.

We find a first warm-up phase of 52 weeks. Appendix J describes the details. We take the first week of which we have appropriate data available and repeat this week 52 times to fill the model. Afterwards, we continue the simulation with the second week. As a second warm-up phase we decide to delete the first 13 weeks of data of the 'actual' simulation run (excluding the first warm-up period). Hence, our second warm-up phase consists of 13 weeks.

This comes down to a run length of 130 weeks, if we use the 78 weeks of available historical data. Of these 130 weeks we use 52 weeks as a first warm-up to fill the model and 13 weeks as a second warm-up to reduce initialisation bias. In the end, we have 65 weeks left for output analysis. Figure 4-5 visualises this set-up and shows by an example how the development of a KPI can look like during this run. The line shows that the simulation model is filled after 52 weeks as the line reaches a steady state. After this phase the simulation starts simulating the other weeks. Our historical data show the growth of Infineon during the last years, which is also simulated by our simulation model.

We determine the number of replications to ensure that our output lies in a 95% confidence interval (α=0.05) and has a relative error of less than 5% (γ=0.05). We need to execute the experiments at least 2 times. As this number is very low, we decide to replicate each of the experiments 10 times. According to our calculations this comes down to a confidence interval of >99.5% and a relative error of <0.5%. The calculations are described in Appendix K.

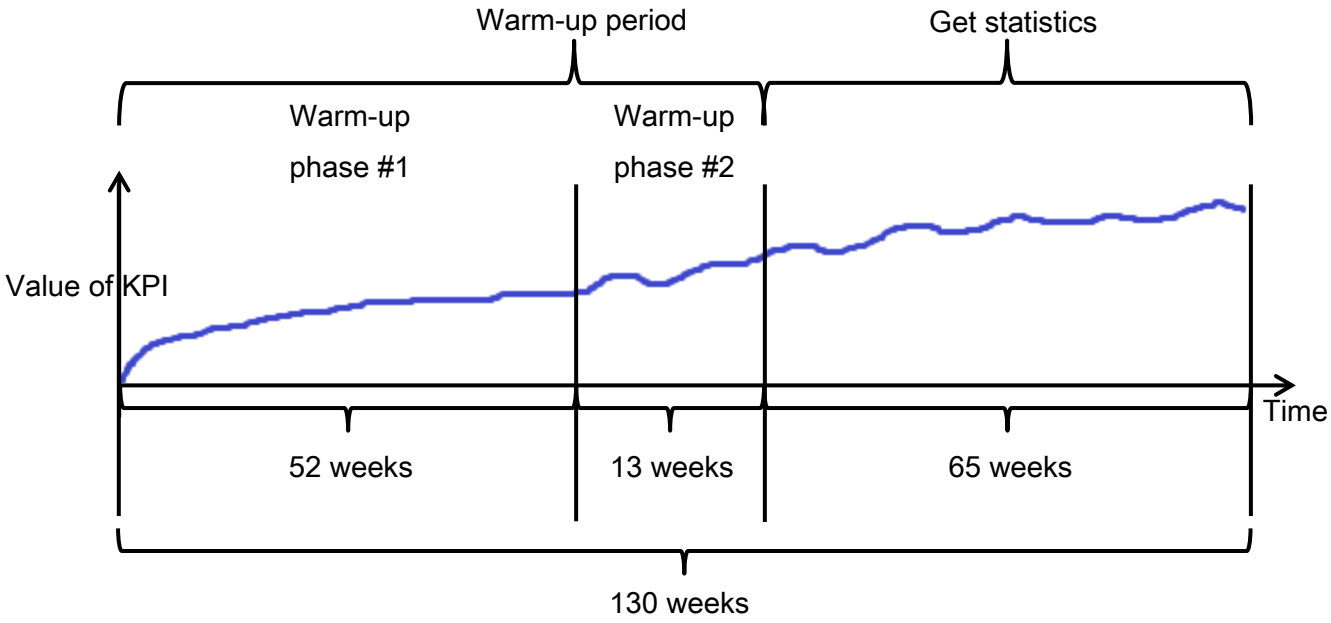


Figure 4-5 Overview of set-up of the run length, with an example of the development of a KPI during this run (blue)

4.2.2 Input data

For the validation experiment of the simulation model of July we use trace driven simulation, for which we use historical data of 78 weeks, from fiscal week 401 until fiscal week 525. This historical data consists of the number of lots started per week and the average lot size per product per week (on material number level, the product number used within Infineon), the variability (α) per process group per week, the number of tools per bottleneck per week, and the availability per bottleneck per week.

4.2.3 Results

By historical data validation we can determine the accuracy of the simulation model by comparing the simulated output with the empirical (historical) data. As stated in Section 1.4 the goal of this research is to achieve a maximal overall average margin between the simulated KPIs data and the historical KPIs data, Δ (see 2.2.4, Equation 8) of 5%. Table 4-2 shows the overall Δs of the averages of the KPIs over all facilities of both the existing and the updated simulation model (see Appendix H for a short description of the KPIs). Similar results are achieved for each facility (Appendix L).

The accuracy of the simulation model has improved, especially when looking at the CT, for which the accuracy measure improves from -21.9% to 1.1%. We explain the differences between the existing and the updated simulation model and the current discrepancies in

more detail in Appendix L. Besides for LOPW, all measures are better than the targeted level of accuracy.

Below we provide explanations for why the simulated LOPW are too high in general and why the accuracy of this measure seems to have worsened (the deviation increased) in comparison to the existing model.

First, we explain the increase in the accuracy measure from 4.4% in the existing model by 1.6% points beyond the target of 5%. The reason is that we completed the input data when updating the model. By completing the input data, we simulate a higher production volume in the updated model as indicated by the improved accuracy (+1.7% points) of the WSPW measure. We see effects in a similar order of magnitude for the WOPW and the LSPW measures as well.

Second, we observe that the output measures show a positive deviation (+6.0% and +3.7% for LOPW and WOPW, respectively). However, the input measures match reality closer (+1.1% and -0.4% for LSPW and WSPW, respectively). This discrepancy can be explained by the fact that we do not simulate wafer losses in both the existing and the updated model. Wafer losses refer to events in which whole wafers are damaged or show such a low yield that they are removed from further processing. In this case, they do not appear in WOPW and LOPW statistics in reality. In both versions of the simulation model (for validation purposes), however, all wafers that appear in the input measures (WSPW, LSPW) will appear in the output measures (WOPW, LOPW) as well.

Both effects together, make the accuracy measure for LOPW violating its targeted level.

Table 4-2 Overall accuracy per KPI; existing versus updated simulation model

Simulation model		WSPW	WOPW	LSPW	LOPW	CT	Utilisation	FF
	Existing	-2.1%	2.5%	-0.8%	4.4%	-21.9%	-3.0%	-4.8%
	Updated	-0.4%	3.7%	1.1%	6.0%	1.1%	-1.0%	2.4%

4.3 Conclusion

In this chapter we discussed the changes we made to the existing simulation model, leading to a more flexible and accurate simulation model. We obtained the flexibility by providing an easy way to change bottlenecks for each of the facilities in case the bottlenecks change. We conclude that we reached the overall accuracy goal (maximal deviation of 5% from the real

data) for all KPIs for the updated simulation model, except for LOPW, which, as explained, does not mean that the simulation model worsened on this point compared to the existing simulation model. It is a side-effect from completing the input data and the fact that wafer losses are not modelled.

5 Balancing inventory and equipment contingencies

In Section 1.6 we formulated the main research question of this research: how can the equipment and inventory contingencies be balanced in a most cost efficient way for a defined α, β, or γ service level (SL^{α}, SL^{β}, and SL^{γ}) using a flexible and accurate supply chain simulation model of Infineon Technologies? By the results of this chapter, which we obtain by multiple experiments in which we vary the decision parameters (die bank reach, DC reach and utilisation) and collect statistics to determine the value of our KPIs (SL^{α}, SL^{β}, and SL^{γ}, and (total) costs), we are able to give an answer to this question.

We first describe the KPIs in Section 5.1, followed by the experimental design in Section 5.2. Section 5.3 informs about the chosen warm-up period, the run length, and number of replications. Section 5.4 explains how we selected the input data. Section 5.5 lists the assumptions made for this experiment and Section 5.6 shows the obtained results of the contingency experiments.

5.1 KPIs

We want to perform a quantitative analysis on the experiments with respect to total costs and the different service level types SL^{α}, SL^{β}, and SL^{γ} as described in Chapter 1 (Section 1.2). These service levels refer to the service levels at the DC. How these KPIs are defined is explained in this section.

For our research, SL^{α} measures the probability that demand can be met completely during a certain time period, without considering backorders from previous weeks. It is an adjusted version of the in literature known non-stock out probability, in which the backorders from previous weeks are considered.

For any product i in a week t, the SL^{α} (SL^{α}_{rit}) is either 0 or 1, dependent on whether the demand of week t (without backorders of the previous weeks) can be fulfilled by the production completely (1) or not (0). Hence, if, for example, only 999 of the 1000 demanded products of a week are delivered from the DC to the customer, the SL^{α} is 0. Aggregating across all products (i=1..I), weeks (t=1..T), and replications (r=1..R), we define SL^{α} as follows:

$$SL^{\alpha} = \frac{\sum_{r=1}^{R}\sum_{i=1}^{I}\sum_{t=1}^{T} SL_{rit}^{\propto}}{R * T * I} \quad (10)$$

The SL^{β} (fill rate) measures the relative backorder level. For a week t and product i the SL^{β} is the fraction of the targeted demand that can be fulfilled during this week ($\mathrm{SL}^{\beta}_{\mathrm{rit}}$). This is based on the backorder change and the demand ($\mathrm{D}_{\mathrm{rit}}$) of week t (Equation 11). The backorder change is calculated by subtracting the total number of backorders of at the end of the previous week ($\mathrm{bo}_{\mathrm{ri,t-1}}$) from the number of backorders at the end of week t ($\mathrm{bo}_{\mathrm{rit}}$).

$$SL^{\beta}_{rit} = max\left\{1 - max\left\{\frac{bo_{rit} - bo_{ri,t-1}}{\mathrm{D}_{\mathrm{rit}}}, 0\right\}, 0\right\} \quad (11)$$

The number of backorders at the end of week t ($\mathrm{bo}_{\mathrm{rit}}$) is determined by the demand ($\mathrm{D}_{\mathrm{ri\omega}}$) until week t and the actual delivered products ($\mathrm{d}_{\mathrm{ri\omega}}$) until week t:

$$bo_{rit} = max\left\{\left(\sum_{\omega=1}^{t} D_{ri\omega} - d_{ri\omega}\right), 0\right\} \quad (12)$$

Aggregating across all products (i=1..I), weeks (t=1..T) and replications (r=1..R), we define SL^{β} as follows:

$$SL^{\beta} = \frac{\sum_{r=1}^{R}\sum_{i=1}^{I}\sum_{t=1}^{T} SL^{\beta}_{rit}}{R * T * I} \quad (13)$$

SL^{γ} (adjusted fill rate) extends the SL^{β} by including a time component and considers backorders of previous weeks as well. Hence, it indicates how fast a production system can recover from backorders. We calculate the SL^{γ} for week t and product i ($\mathrm{SL}^{\gamma}_{\mathrm{rit}}$) by:

$$SL^{\gamma}_{rit} = max\left\{1 - \frac{bo_{rit}}{D_{rit}}, 0\right\} \quad (14)$$

Aggregating across all products (i=1..I), weeks (t=1..T) and replications (r=1..R), we define SL^{γ} as follows:

$$SL^{\gamma} = \frac{\sum_{r=1}^{R}\sum_{i=1}^{I}\sum_{t=1}^{T} SL^{\gamma}_{rit}}{R * T * I} \quad (15)$$

For simplification, we use average cost estimates for the financial analysis of the results. These estimates are not product-specific. Aggregating across all products (i=1..I), weeks (t=1..T) and replications (r=1..R), we determine the average total costs (TC) by:

- the stock level of the die bank ($\mathrm{s}^{\mathrm{db}}_{\mathrm{rit}}$) at the end of week t multiplied with the holding costs at the die bank ($\mathrm{hc}^{\mathrm{db}}$)
- the stock level of the DC at the end of week t ($\mathrm{s}^{\mathrm{dc}}_{\mathrm{rit}}$) multiplied with the holding costs at the DC ($\mathrm{hc}^{\mathrm{dc}}$)

- the utilisation (UUm), as defined by the experiments, multiplied with the capacity costs (cc)
- the WIP in the FE at the end of week t ($\mathrm{WIP}_{\mathrm{rit}}^{\mathrm{FE}}$) multiplied with the FE WIP costs ($\mathrm{wc}^{\mathrm{FE}}$)
- the WIP in the BE at the end of week t ($\mathrm{WIP}_{\mathrm{rit}}^{\mathrm{BE}}$) multiplied with the BE WIP costs ($\mathrm{wc}^{\mathrm{BE}}$).

This can be formulated as follows:

$$TC = \Big(\sum_{r=1}^{R}\sum_{i=1}^{I}\sum_{t=1}^{T}\big(s_{rit}^{db} * hc^{db} + s_{rit}^{dc} * hc^{dc} + UUm * cc + \mathrm{WIP}_{rit}^{FE} * \mathrm{wc}^{FE} + \mathrm{WIP}_{rit}^{BE} * \mathrm{wc}^{BE}\big)\Big) / (R * T) \quad (16)$$

All resources costs are included in our formulation. We decide to exclude the stock out costs from our total costs calculation, because these costs are related to the service levels. As we want to integrate service levels as a minimum requirement into our analysis, we want to keep the service levels and the total costs separated and independent from each other to make the factor level combinations (see Section 5.2) comparable. The resources costs are not related to the customer satisfaction (and therefore service level) and cause therefore no problem for our analysis.

Due to confidentiality reasons we scale the costs for this thesis. The costs for the base case are the most expensive and are therefore translated to costs of 100%. The financial values are obtained based on the opinion of the financial expert of the Scenario and Flexibility Planning team and the current practices at Infineon (Stang, 2015). The relation between the WIP and holding costs is rated as follows: $0.20\mathrm{wc}^{\mathrm{FE}}$ equals $0.39\mathrm{hc}^{\mathrm{db}}$ equals $0.64\mathrm{wc}^{\mathrm{BE}}$ equals $1\mathrm{hc}^{\mathrm{dc}}$. The capacity costs are based on the investment costs and a depreciation factor of 5 years.

5.2 Experimental design

Running a wafer fab at a high utilisation to keep the return on investment high increases the manufacturing CT, which might require increased inventory levels to stay reactive towards demand variations on short notice. Therefore, Infineon aims at finding a balance between inventory and equipment contingencies, so that they are able to deliver on time to their customers against minimal costs.

For the experiments to balance inventory and equipment contingencies we use the production plan function (Appendix D) of the model. This function uses a demand scenario (production demand) to calculate the number of lots to be released based on yield, WIP, stock levels, planned CT, and target reaches. These lots are released during the week following the call of the plan function with constant inter arrival times. The execution of the production plan results in a certain utilisation of the equipment. Based on this logic, shown in Figure 5-1, the experiments are designed.

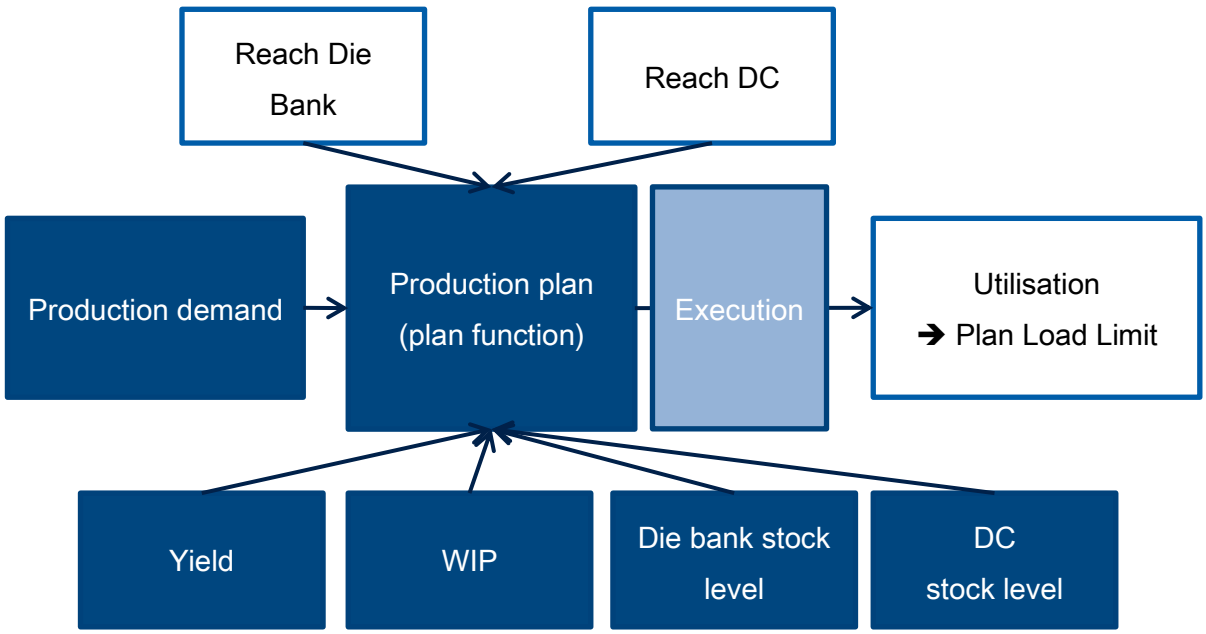


Figure 5-1 Relationships between production demand, the plan function and the utilisation

We define a base case, with a low utilisation of 75% of the steppers and a die bank and DC reach of 35 days. We expect with this low utilisation that the lead times will be short, which makes it easier to replenish the stocks, and that because of high stock levels the service level types are expected to be close to 100%. We note that this is the most expensive tested factor level combination, because utilisation is the lowest (unused capacity is the highest) and stock levels are the highest.

We decide to start with a utilisation of 75% as this is the lowest value that is usually used as Plan Load Limit (see Section 2.1.2). We chose a die bank and DC reach of 35 days as upper limit, because a reach of 28 days is typically used by the supply chain planner for the inventory planning at Infineon. As this is based mainly on experience, we decide to study a higher reach of 35 weeks as well.

We increase the utilisation (until 95% with an increment of 5%) of the stepper bottlenecks by using dummy products and we assume that the simulated utilisation (in combination with the

inventory contingency) corresponds with the fraction of the uptime that Infineon should use for the production plan: the Plan Load Limit. We only increase the utilisation of the steppers, because lithography equipment is the typical bottleneck in the semiconductor industry.

The reach is varied at the die bank and the DC; from 0 till 35 days, with an increment of 7 days. By balancing the inventory levels with equipment capacity we expect that more stock is needed. Table 5-1 summarises the factors and factor levels we want to experiment with. We make 5 replications of each experiment. How this number is determined was explained in Section 4.3.2.

Table 5-1 Experimental design; bold the base case

Factor	Levels	# levels	# factor level combinations
Utilisation (%)	{**75%**, 80%, 85%, 90%, 95%}	5	180
Reach Die Bank (days)	{0, 7, 14, 21, 28, **35**}	6	
Reach DC (days)	{0, 7, 14, 21, 28, **35**}	6	
Number of simulation runs (5 replications per experiment):			900

5.3 Warm-up period, run length, and number of replications

We determine the warm-up period as described in Section 4.2.1. Based on the average utilisation of the steppers per week, we find a warm-up period of 52 weeks. We use the utilisation as a base, because it has to reach a steady state at 75%, 80%, 85%, 90% or 95% to represent the utilisation factor levels. We determine the warm-up period for the lowest and highest utilisation, on average 75% and 95%, respectively, over all stepper work centres and facilities, with a die bank and DC reach of 0 and 35 days. We assume these factor combinations to represent the upper and lower limit of the warm-up period. With a low utilisation we expect that the products can be processed smoothly which requires a low reach. This provides us the lower limit case. With a very high utilisation we expect that the production lead times will increase, which will require a high reach as described in Section 1.3. This provides us the upper limit case.

We found a lower limit of 38 weeks (Villach, 75%, die bank and DC reach 0 days) and an upper limit of 52 weeks (Kulim, 95%, die bank and DC reach 35 days). Therefore we use the upper limit of 52 weeks for all experiments for better comparability.

As we want to simulate one year, we set the run length to 104 weeks. Considering the warm-up period, we only collect statistics between week 53 and 104 and base our results on this data.

In Section 4.3.1 we wrote that we decided to replicate the contingency experiments 5 times. This number is established by analysing the average UUm in the fab per week per replication by the same procedure we used for the determination of the number of replications for the validation experiment as described in Appendix K. To ensure a confidence level of at least 99%, with a relative error of at most 1%, we need to make at least 4 replications.

5.4 Input data

For our contingency experiments we use a stable demand pattern for 104 weeks. This demand is based on the historical data of fiscal week 525. Of this data we take per facility the three process groups with the highest production volume, based on the number of layers (lithography steps), and search for each process group a product on material number level (the product number used within Infineon). We take the average WSPW of each process group as an input starting point and scale the WSPW to reach the theoretical utilisation of the base case: 75%. This demand scenario is added to the input file and is read into the plan function (Appendix D). The plan function uses, among other factors, the planned CT per product. This planned CT is adjusted in reality when the manufacturing CT changes significantly as the manufacturing CT strongly depends on utilisation. Therefore, we adjust the planned CT for all utilisation levels for experimentation as well. We determine this planned CT by running the model one time (1 replication), with a die bank and DC reach of 35 days, for each theoretical utilisation level and by taking the average CT per product over the last 52 simulated weeks.

5.5 Assumptions

The assumptions we made for the contingency experiments are:

- All demand is committed by the customers. Demand is a production target, rather than customer demand, because there are no customers modelled.
- Pre-delivery is not possible.
- The different locations of the die banks and the DCs are aggregated.

- We only consider stepper bottlenecks for the capacity contingencies.
- We have unlimited inventory storage space.

5.6 Results

We divided the analysis of the results in two parts: coherence and optimising equipment and inventory contingencies.

Coherence

The results reveal several main relationships between the tested factors, service levels, and costs. Appendix M (Table A-9) lists the detailed results.

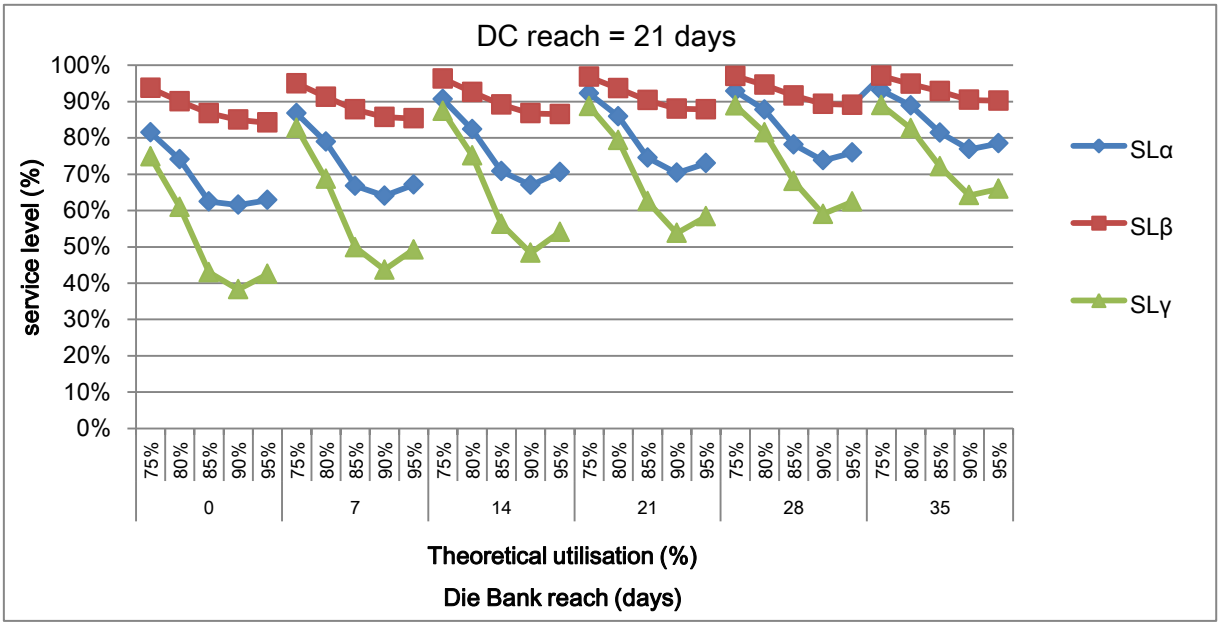


Figure 5-2 Relationship between the different service levels, utilisation and die bank reach

As shown in Figure 5-2 with an example of a fixed DC reach of 21 days, the different service level types have the same pattern; a low utilisation and a high die bank leads to a high service level. The graph looks the same for all DC reaches (see Appendix M, Figures A-8, A-9 & A-10). Therefore, we conclude that also a high DC reach, combined with a low utilisation, leads to a high service level. As the lines for the different service level types have the same pattern (see Figure 5-2), we base further relationship figures only on SL^{β}.

Figure 5-2 also shows the relationship between the utilisation and the different service levels. As the utilisation increases, SL^{α}, SL^{β}, and SL^{γ} typically decrease, because production lead times increase (as described in Section 1.3). This will harm the customer service.

SL^{β} and SL^{γ} increase when the utilisation increases from 90% to 95%. This is not the case for SL^{α}. We expect that the cause lies in the level of backorders which are included in SL^{β}

and SL^{γ}, but not in the SL^{α}. Therefore, we analyse the backorder levels and notice that the backorder level at 95% utilisation is much lower than the backorder level at 90% utilisation (see Appendix M, Figure A-11). An explanation for this lower backorder level is that there are enough buffers. From the stock level data obtained by the experiments we see that with 95% utilisation die bank stock levels are lower, but DC stock levels are higher than with 90% utilisation (see Appendix M, Figure A-12). This explains the lower backorder level and the higher SL^{β} and SL^{γ}, as the service levels are related to the DC.

For a given die bank reach, the SL^{α}, SL^{β}, and SL^{γ} increase with an increasing DC reach and with decreasing utilisation. An example, with a fixed die bank reach of 35 days, is illustrated in Figure 5-3 (left).

Remarkable is that a utilisation of 90% or 95% does not make much difference for the SL^{β}. We have two possible explanations. The first possible explanation for the small difference for the SL^{β} at 90% and 95% utilisation is that we reached the point that we have enough buffers, so it does not make sense to have more stock, as is also brought forward before. Another explanation is that it is caused by the fact that the simulated utilisation and the defined utilisation are not equal (see Table 5-2). Table 5-2 shows that Regensburg and Villach are not able to reach this utilisation on average in the simulation model, we suspect that this is caused by the fact that the plan function does not include production plan smoothing as discussed in Section 2.2.1. Unfortunately, we cannot determine the cause of this utilisation difference by the gathered data and since we are not able to simulate the defined utilisation we are not able to make a statement about the real cause.

Table 5-2 Defined and simulated utilisation for Villach, Regensburg, and Kulim of the steppers

Facility	Defined utilisation	Simulated utilisation
Villach	90%	86%
	95%	86%
Regensburg	90%	85%
	95%	88%
Kulim	90%	90%
	95%	95%

For a given DC reach, the SL^{α}, SL^{β}, and SL^{γ} increase with an increasing die bank reach and with decreasing utilisation. For a DC reach of 35 days, the SL^{α}, SL^{β}, and SL^{γ} are 100% at 75% utilisation, regardless the die bank reach. Even with a DC reach of 28 days, a SL^{β} and

SLᵞ can be reached of 99.8% at 75% utilisation, regardless the die bank reach. Figure 5-3 (right) illustrates these relationships for a fixed DC reach of 35 days.

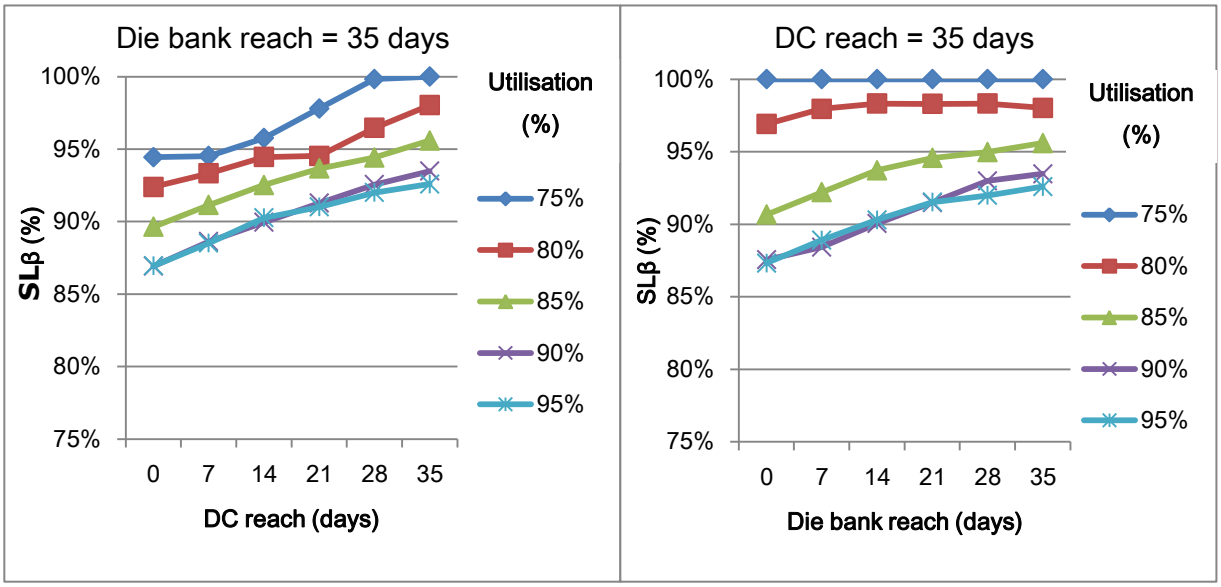


Figure 5-3 Relationship between the SLβ and certain factor level combinations

From a service level perspective, the DC reach is preferable to the die bank reach. Figure 5-3 shows that the DC reach has a higher impact on the SLᵅ, SLᵝ, and SLᵞ than the die bank reach, e.g. at a high DC reach (35 days) the service level is 100% at 75% utilisation, regardless the die bank reach. A die bank reach of 35 days alone does not guarantee a service level of 100%.

Table 5-3 shows the SLᵝ for the different reach factor level combinations (die bank reach and DC reach) and 75% utilisation. By comparing the increase in service level when going to the left in the table with the increase when going down in the table, we can confirm the higher impact of the DC reach for all factor level combinations, except for one; a die bank reach increase from 7 to 14 has a larger impact than a DC reach increase from 0 to 7. Although the DC reach has a larger impact on the different service levels than the die bank reach, from a cost perspective, inventory at the DC is 2.6 times more expensive than die bank inventory (Section 4.3.2). We elaborate on this trade-off later in this section.

Table 5-3 SL^{β} at the different (die bank and DC) reach combinations (at 75% utilisation)

Die bank reach (days) \ DC reach (days)	0	7	14	21	28	35
0	86,6%	87,1%	91,6%	97,2%	99,8%	100,0%
7	87,2%	90,5%	94,7%	97,7%	99,8%	100,0%
14	91,1%	93,5%	95,7%	97,8%	99,8%	100,0%
21	93,2%	94,4%	95,7%	97,8%	99,8%	100,0%
28	94,3%	94,6%	95,6%	97,8%	99,8%	100,0%
35	94,4%	94,5%	95,8%	97,8%	99,8%	100,0%

Figure 5-4 displays the costs related to the tested factor level combinations at a fixed DC reach of 0 days. Although in this figure only the curve for which the DC reach is 0 is shown, the other curves of the other DC reaches overlap these lines a lot and are very similar. Therefore, this figure is a good representation of the relationships between the costs and the factor levels. This figure shows that the utilisation has a large impact on the costs; a low utilisation leads to higher costs than a high utilisation. The die bank and DC reach on the other hand, only have a minor impact on the costs.

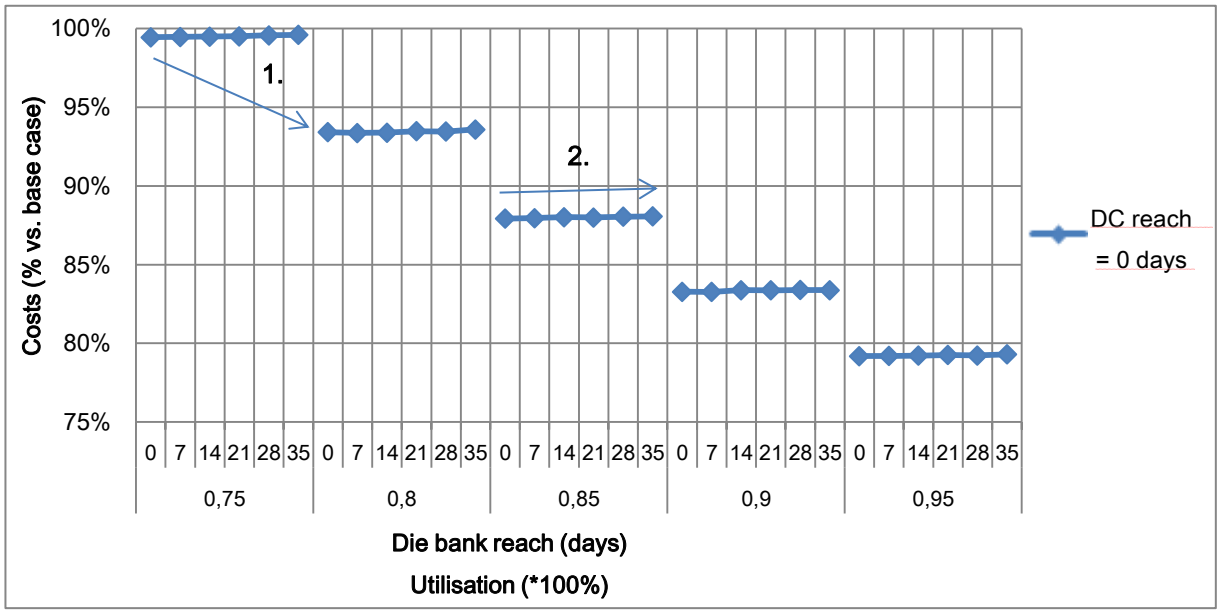


Figure 5-4 Relationship between the costs and the different factor level combinations

1. Impact of the utilisation on the costs (at a die bank reach of 0)
2. Impact of the die bank reach on the costs (at 85% utilisation)

This is what we expected, as the investment costs (capacity costs) are much higher than the inventory costs. As the utilisation increases, the cost difference for the die bank and DC

reach combinations becomes smaller (the lines become more horizontal). For a given DC reach, we see that changing the utilisation has a higher cost impact than changing the die bank reach, because the inventory costs are small compared to capacity costs, which makes the relative cost difference larger.

Optimising the trade-off between inventory and equipment contingencies

The trade-off between inventory and equipment contingencies consists of choosing between certain SL^{α}, SL^{β}, and SL^{γ} and costs. Table A-9 in Appendix M lists the values of these KPIs of the experiments for each factor level combination. As Figure 5-5 shows, lower costs do not always mean a lower service level, because the reaches have a high impact on the service levels (almost vertical red lines) and are relatively cheap compared to the capacity costs (almost horizontal blue lines). These capacity costs contribute most to the total costs. The figure also confirms our previous findings that with a low utilisation and a high DC reach a high service level can be reached.

Table 5-3 shows the main results for different targeted SL^{α}, SL^{β}, or SL^{γ}. For each target the factor level combination resulting in the lowest costs is displayed. To give an example: assume Infineon wants to achieve a SL^{β} of at least 95% at the lowest costs, then a die bank and DC reach of 35 days and a utilisation of 85% has to be chosen to plan production.

As the table shows, to achieve a certain SL^{β} at lowest costs, having a high die bank reach and utilisation is more important than having a high DC reach. We explain this by the fact that SL^{β} does not include backorders of previous weeks, so only the demand that could not be fulfilled this week. On the contrary, if we want to achieve a certain SL^{γ}, which also considers backorders of the previous weeks, at lowest cost, it is more important to have a high DC reach instead of a high die bank reach. By keeping the inventory at stocking points close to the customer, a better service level SL^{γ} can be reached, because reaction time is the fastest there. We can also keep the utilisation low to reach a certain SL^{γ}. To achieve certain SL^{α}, both a high die bank reach and a high DC reach are important.

The costs for aiming at a certain SL^{γ} are higher than the costs for aiming at a certain SL^{α} or SL^{β}. This can be explained by the fact that this service level measures how fast a company can recover from backorders and includes a time component the other two service level types have not. This service level type (SL^{γ}) is lower by definition, as described in Section 5.1 and shown in Figure 5-2, so it is harder to reach the same service level as SL^{α} or SL^{β}.

Table 5-4 Main results of optimising the trade-off between inventory and equipment contingencies

		How to achieve target at lowest costs			KPI values			
	Target	Reach die bank (days)	Reach DC (days)	Utilisation (%)	SLα (%)	SLβ (%)	SLγ (%)	Costs (%) vs. base case
SLα (%)	>95%	14	35	80%	96,2%	98,3%	96,6%	93,8%
	>90%	7	28	80%	90,9%	95,4%	89,5%	93,7%
	>85%	35	35	95%	85,3%	92,6%	74,4%	79,6%
	>80%	35	21	95%	81,3%	91,0%	69,9%	79,5%
SLβ (%)	>95%	35	35	85%	88,6%	95,6%	83,6%	88,5%
	>90%	35	14	95%	79,2%	90,3%	66,6%	79,4%
	>85%	28	0	95%	64,8%	85,9%	48,1%	79,2%
	>80%	0	0	95%	49,7%	81,6%	18,5%	79,2%
SLγ (%)	>95%	14	35	80%	96,2%	98,3%	96,6%	93,8%
	>90%	0	35	80%	93,4%	96,9%	93,6%	93,7%
	>85%	0	28	80%	87,3%	94,6%	85,5%	93,6%
	>80%	21	35	95%	86,4%	94,6%	81,8%	88,3%

5.1 Conclusion

In this chapter we discussed the experiments to balance inventory and equipment contingencies. We tested 180 factor level combinations to balance inventory and equipment contingencies. We used the SL^{α}, SL^{β}, and SL^{γ} and the total costs as KPIs to assess how we can achieve a targeted service level at lowest costs. We conclude that producing at a low utilisation has a big negative impact on the costs although this positively influences the service level. From a service level perspective, the DC reach is preferable to the die bank reach. We think this is a legitimate result, as the company can react faster to incoming orders and is also still able to deliver in case of disturbances in the supply chain. The die bank provides these benefits as well, but to a lower degree. However, it would be cheaper to place stock at the die bank. Therefore, there exists another trade-off which is out of scope in thesis, but should be investigated further. Furthermore, we find that for each service level type different factor level combinations lead to the lowest costs.

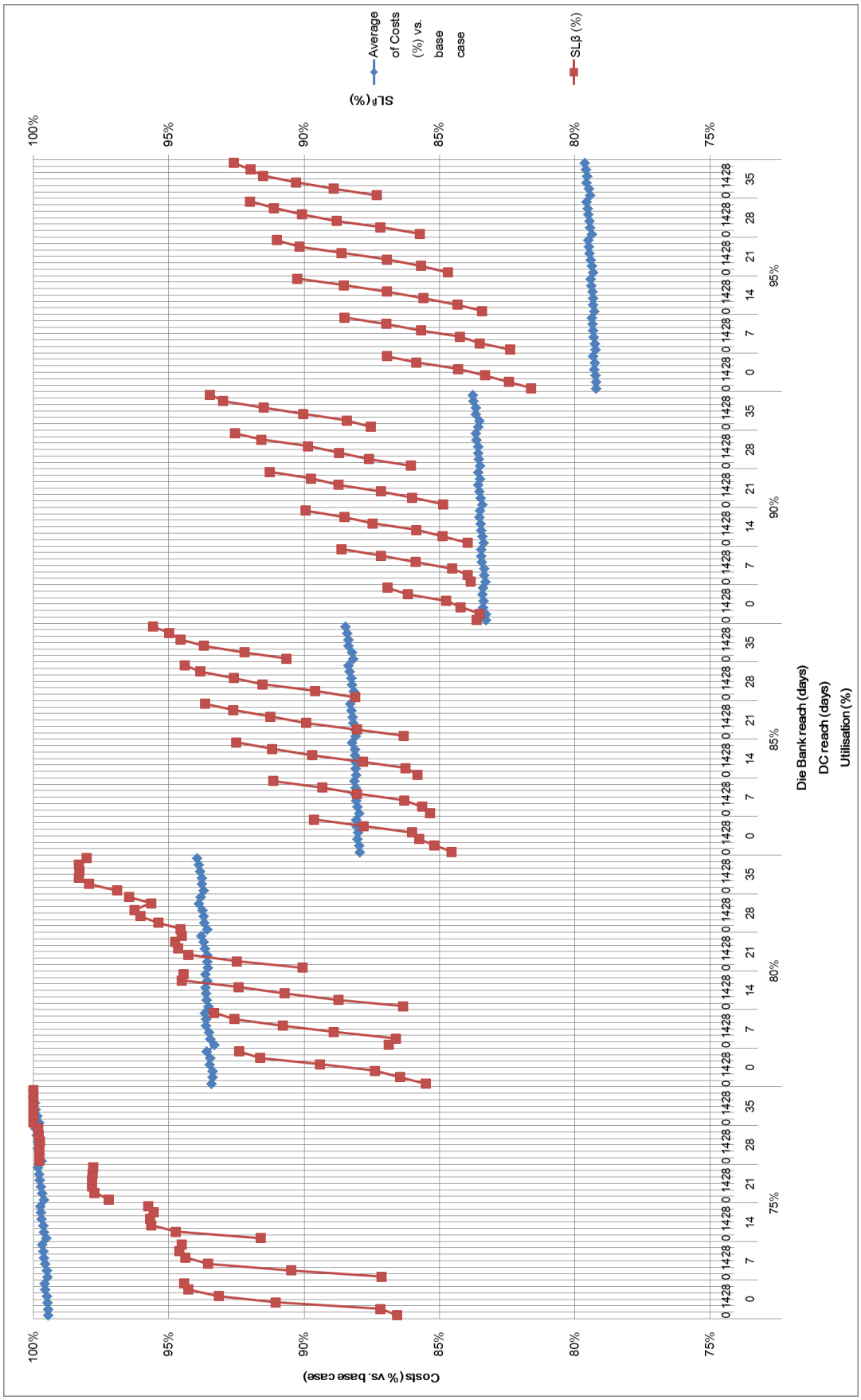


Figure 5-5 Relationship between the costs, SL$^{\beta}$, and the factor level combinations

6 Conclusion and recommendations

Section 5.1 contains the conclusions of this research and discusses its limitations. In Section 5.2 we give directions for further research and do several recommendations.

6.1 Conclusion and discussion

This research consisted of two parts: balancing inventory and equipment contingencies and adjusting the supply chain simulation model with which this balance can be found. We formulated the following main research question:

How can the equipment and inventory contingencies be balanced in a most cost efficient way for a defined α, β, or γ service level using a flexible and an accurate supply chain simulation model of Infineon Technologies?

We found several existing approaches in the literature for supply chain simulation in general, discussing reusability and the level of detail of modelling, and simulation approaches for modelling a wafer fab. We considered these approaches for the supply chain simulation model of Infineon.

We developed a flexible supply chain simulation model in which the bottlenecks can be adjusted easily. With this model we conducted a validation experiment and compared it with empirical data. From this experiment we concluded that the simulation model is accurate enough for our following experiments on balancing inventory and equipment contingencies, with a maximal deviation of 5% from the real data. Especially the accuracy of the CT improved from -21.9% to 1.1%. The ability to model a wafer fab by using simplifications as discussed in literature has proven to represent reality in a simplified yet accurate way.

In our experiments we varied the die bank and DC reach to represent the inventory contingency and the utilisation to represent the equipment contingency. For each simulated factor level combination we collected the following KPIs: SL^{α}, SL^{β}, and SL^{γ} and the total costs.

We used a simplified demand scenario, as our goal was to understand and quantify the relationships. The observed relationships were plausible. We conclude that from a financial point of view, a high utilisation is beneficial compared to high inventory levels. However, from a service level point of view, regardless the service level type (SL^{α}, SL^{β}, and SL^{γ}), a high inventory level is more of use, especially the inventory at stocking points close to the customer, because reaction time is the fastest there. DC stock is, however, from a cost perspective less preferable than die bank stock, as the DC stock is more expensive (in the

considered case by a factor of 2.6) than the die bank stock because of the diversification of the product at the stock points and the added value at the BE.

It appeared difficult for the simulation model to simulate a high utilisation of the bottlenecks, as difference in defined and simulated utilisation showed. We assume this is caused by the fact that the current plan function incorporates no production smoothing, such as scheduling the demand earlier for production if capacity in a later period is not sufficient. This makes, however, the obtained results for utilisations of 90% and 95% less legitimate, although this is the utilisation you want to be at as a company from a cost point of view.

We have to note that the SL^{α}, SL^{β}, and SL^{γ} will be higher in reality, because in the simulation model dispatching is modelled as FIFO, but in reality this prioritisation is done based on Earliest Operations Due Date. Moreover, measures exist to prioritise lots that are urgently needed. Furthermore, we only used a limited number of products as input for our contingency experiments. This can lead to different results than when the full product mix is used.

Despite the limitations of this research, the results give a good overview of the relationships between inventory and equipment contingencies and the related trade-offs.

6.2 Recommendations

We recommend utilising the available equipment as high as possible and buffer at the distribution centres and die banks to make up for being less flexible and slower due to an increased flow factor. What “as high as possible” means needs to be studied with more detailed scenarios and a more detailed simulation model. Therefore, we give both contingency-related and model-related recommendations in this section.

We discuss the recommendations for future research in different categories: simplification, reusability, plan, make, contingency experiments, data, and inventory contingency.

Simplification

We suggest to compare the current simplified model to a more detailed model. This is also seen in literature, but mostly no historical data is used as a reference point to measure accuracy performance. We think this could extend the existing literature and prove whether or not and to which extend a detailed model’s performance is better.

Reusability

We propose to increase the reusability of the simulation model. Some code lacks background information about the decisions made, so that others can follow the reasoning behind the code. More extensive comments in the simulation code increase reusability. We are of the opinion that at this moment, it takes a lot of time to get to know and understand the whole model.

Plan

We recommend to replace the simplified planning function by a more advanced object. Two former graduate students at Infineon created simulation objects that include plan stability and represent the demand-supply matching of Infineon to a high degree (Guo, 2015; Würf, 2015). These objects also incorporate ways to smoothen the production plan. How and to which extend this can be incorporated in the supply chain simulation model needs to be researched.

Make

We propose to include a more detailed BE and to extend the FE object in the simulation model. The BE facilities can be modelled with their own bottlenecks, following the same approach as the modelling approach for the wafer fab object. In the FE, we think that the equipment representation can be improved by, for instance, incorporating cluster tools. This makes the use of a capacity factor redundant. Apart from that, the inclusion of more bottlenecks can be tested. Furthermore, we think the simulation can be extended, for example by including the possibility to study other wafer diameters. The 300mm wafer facility in Dresden should also be modelled for this purpose. Moreover, we think that the possibility to change the queue prioritisation to Earliest Operations Due Date should be researched to gain more reliable service levels and hence, more reliable experimental results.

Contingency experiments

We recommend to use a more realistic demand scenario with a broader product mix and level demand uncertainty and to use a lower increment when increasing the factor levels (e.g. the utilisation or the reaches). By this end, more specific answers can be obtained when conducting contingency experiments.

Data

We suggest to document the way data in the databases are obtained (when, where, and how) and to set-up databases that contain data specifically for simulation purposes. The next step is to link the simulation model to the databases that are tailored to simulation. As simulation is an upcoming scenario analysis tool at Infineon, we expect this to increase the accuracy of the simulation models. By this end, the use of scaling factors is not needed which increases the credibility of the simulation model for people outside the department.

Inventory contingency planning

We propose to experiment with other inventory planning approaches. Even if there is a simulation model available, simulation studies take a long time for parameterisation, running the simulation model and analysing the outputs. Therefore, we need to make a pre-selection of the factor levels we want to test. In this context, analytical inventory planning approaches (such as models making use of echelon inventory policies (e.g., Van der Heijden, 2014; Simchi-Levi, Kaminsky, & Simchi-Levi, 2008)) can help to narrow down the solution space to a few relevant scenarios that are a starting point for analysis with a simulation model, such as the one developed in this thesis.

References

Álvarez Luque, M. (2015). *Modelling and evaluation of stock target setting approaches in a semiconductor supply chain using agent-based simulation.* Hamburg: Technische Universität Hamburg-Harburg.

AnyLogic. (2015, March). Retrieved from AnyLogic: http://www.anylogic.com/downloads

Atamatürk, A., & Hochbaum, D. (2001). Capacity acquisition, subcontracting, and lot sizing. *Management Science, 47*(8), 1081-1100.

Aurand, S., & Miller, P. (1997). The operating curve: a method to measure and benchmark manufacturing line performance. *Advanced Semiconductor Manufacturing Conference and Workshop, 1997. IEEE/SEMI* (pp. 391-397). Cambridge, MA : IEEE.

Axsäter, S. (2006). *Inventory control* (2nd ed.). New York: Springer.

Balanchandran, K., Li, S., & Radhakrishnan, S. (2007). A framework for unused capacity: theory and empirical analysis. *The Journal of Applied Management Accounting Research, 5*(1), 21-37.

Banks, J. (1998). *Handbook of simulation: principles, methodology, advances, applications, and practice.* New York, NY: John Wiley & Sons Ltd.

Borshchev, A. (2013). *The big book of simulation modeling: multimethod modeling with AnyLogic 6.* AnyLogic North America.

Bradley, J., & Arntzen, B. (1999). The simultaneous planning of production, capacity and inventory in seasonal demand environments. *Operations Research, 47*(6), 795-806.

Brandon-Jones, A., & Slack, N. (2008). *Quantitive analysis in operations management.* Essex: Pearson Education Limited.

Brooks, R. J., & Tobias, A. M. (2000). Simplification in the simulation of manufacturing systems. *International Journal of Production Research, 38*(5), 1009-1027.

Brown, A. O., Lee, H. L., & Petrakian, R. (2000). Xilinx improves its semiconductor supply chain using product and process postponement. *Interfaces, 30*(4), 65-85.

Chance, F., Robinson, J., & Fowler, J. (1996). Supporting manufacturing with simulation: model design, development, and deployment. *Proceedings of the 1996 Winter Simulation Conference* (pp. 114-121). San Diego, CA: IEEE.

Chien, C.-F., Dauzère-Pérès, S., Ehm, H., Fowler, J., Jiang, Z., & Krishnaswamy, S. (2011). Modelling and analysis of semiconductor manufacturing in a shrinking world: challenges and successes. *European Journal of Industrial Engineering, 5*(3), 254-271.

Duarte, B., Fowler, J., Knutson, K., Gel, E., & Shunk, D. (2002). Parameterization of fast and accurate simulations for complex supply networks. *Proceedings of the 2002 Winter Simulation Conference* (pp. 1327-1336). San Diego, CA: IEEE.

Ehm, H. (2015, January 15). Guest Lecture "Managing complex supply chains in the high-tech industry and challenges in the planning process". Munich, Germany.

Ehm, H., Ponsignon, T., & Kaufmann, T. (2011). The global supply chain is our new fab: integration and automation challenges. *Advanced Semiconductor Manufacturing Conference (ASMC)* (pp. 1-6). Saratoga Springs, NY: IEEE.

Eirich, S. (2014). *Improving operational production scheduling: optimal production priority setting in semiconductor manufacturing.* Munich: Technische Universität München.

Fowler, J. W., & Rose, O. (2004). Grand challenges in modeling and simulation of complex manufacturing systems. *Simulation, 80*(9), 496-476.

Gallego, G., Katircioglu, K., & Ramachandran, B. (2006). Semiconductor inventory management with multiple grade parts and downgrading. *Production Planning & Control, 17*(7), 689-700.

Guo, J. (2015). *Measuring and reducing plan instability: a case of a semiconductor integrated device manufacturing company.* Madrid: Universidad Politécnica de Madrid.

Gupta, J., Ruiz, R., Fowler, J., & Mason, S. (2006). Operational planning and control of semiconductor wafer production. *Production Planning & Control, 17*(7), 639-647.

Hoad, K., Robinson, S., & Davies, R. (2008). Automating warm-up length estimation. *Proceedings of the 2008 Winter Simulation Conference* (pp. 532-540). Austin, TX : IEEE.

Hopp, W., & Spearman, M. (2008). *Factory physics.* New York, NY, USA: McGraw-Hill.

Hung, Y.-F., & Leachman, R. (1999). Reduced simulation models of wafer fabrication facilities. *International Journal of Production Research, 37*(12), 2685-2701.

Infineon Technologies. (1998). *Technical regulation 26 - overall equipment effectiveness.* Munich.

Infineon Technologies. (2008). *Technical regulation 25 - manufacturing equipment performance: reliability, maintenance, maintainability.* Munich.

Infineon Technologies. (2010). *Technical regulation 27 - production performance indicators.* Munich.

Infineon Technologies AG. (2014). *Annual Report 2014: Systematic growth.* Neubiberg: Infineon Technologies AG.

Jain, S., Lim, C.-C., Gan, B.-P., & Low, Y.-H. (1999). Criticality of detailed modeling in semiconductor supply chain simulation. *Proceedings of the 1999 Winter Simulation Conference* (pp. 888-896). New York, NY: ACM.

Kleijnen, J. (1995). Verification and validation of simulation models. *European Journal of Operational Research*(82), 145-162.

Law, A. M. (2007). *Simulation modelling analysis* (4th ed.). Tucson, Arizona, USA: McGraw-Hill.

Minner, S. (2012). *Inventory Management.* Munich: Technische Universität München.

Morrice, D., & Valdez, R. (2005). Discrete event simulation in supply chain planinng and inventory control at Freescale semiconductur, Inc. *Proceedings of the 2005 Winter Simulation Conference* (pp. 1718-1724). Orlando, FL: IEEE.

Oechsner, R., Pfeffer, M., Pfitzner, L., Binder, H., Müller, E., & Vonderstrass, T. (2003). From overall equipment efficiency (OEE) to overall fab effectiveness (OFE). *Materials Science in Semiconductor Processing*(5), 333-339.

Peikert, A., Thoma, J., & Brown, S. (1998). A rapid modeling technique for measurable improvements in factory performance. *Proceedings of the 1998 Winter Simulation Conference* (pp. 1011-1015). Piscataway, NJ: IEEE.

Piplani, R., & Puah, S. (2004). Simplification strategies for simulation models of semiconductor facilities. *Journal of Manufacturing Technology Management, 15*(7), 618-625.

Robinson, S., Nance, R., Paul, R., Pidd, M., & Taylor, S. (2004). Simulation model reuse: definitions, benefits and obstacles. *Simulation Modelling Practice and Theory, 12*, 479-494.

Rose, O. (2004). Modeling tool failures in semiconductor fab simulation. *Proceedings of the 2004 Winter Simulation Conference* (pp. 1910-1914). Washington, DC: IEEE.

Rose, O. (2007). Improved simple simulation models for semiconductor wafer factories. *Proceedings of the 2007 Winter Simulation Conference* (pp. 1708-1712). Washington, DC: IEEE.

Sargent, R. (2013). Verification and validation of simulation models. *Journal of Simulation*(7), 12-24.

Shannon, R. E. (1988). Introduction to the art and science of simulation. *Proceedings of the 1988 Winter Simulation Conference* (pp. 7-14). Washington, DC: IEEE.

Simchi-Levi, P., Kaminsky, P., & Simchi-Levi, E. (2008). *Designing and managing the supply chain: concepts, strategies, and case studies* (3rd ed.). New York: McGraw-Hill/Irwin.

Slack, N., Chambers, S., & Johnston, R. (2007). *Operations management* (5th ed.). Essex, England: Pearson Education Limited.

Smith, C., Minor, E., & Jen, H. (1995). Regression-based due date assignment rules for improved assembly shop performance. *International Journal of Production Research, 33*(9), 2375-2385.

Sprenger, R., & Rose, O. (2011). On the simplification of semiconductor wafer factory simulation models. In S. Robinson, R. Brooks, K. Kotiadis, & D. Van der Zee, *Conceptual modelling for discrete-event simulation* (pp. 451-470). Boca Raton, FL: CRC Press, Taylor and Francis Group.

Stang, M. (2015, August 10). Manager Invest Consolidation. (A. Adriaansen, Interviewer)

Terzi, S., & Cavalieri, S. (2004). Simulation in the supply chain context: a survey. *Computers in Industry*(53), 3-16.

Uszoy, R., Lee, C.-Y., & Martin-Vega, L. A. (1992). A review of production planning and scheduling models in the semiconductor industry part I: system characteristics, performance evaluation and production planning. *IEE Transactions, 24*(4), 47-60.

Van der Heijden, M. (2014). *Inventory allocation in distribution networks.* Course materials, Enschede.

Van der Zee, D., & Van der Vorst, J. (2005). A modeling framework for supply chain simulation: opportunities for improved decision making. *Decision Sciences, 36*(1), 65-95.

Van Mieghem, J., & Rudi, N. (2002). Newsvendor networks: inventory management and capacity investment with discretionary activities. *Manufacturing & Service Operations Management, 4*(4), 313-335.

Wu, S., Erkoc, M., & Karabuk, S. (2005). Managing capacity in the high-tech industry: a review of literature. *The Engineering Economist, 50*, 125-158.

Würf, C. (2015). *Impact analysis of capacity modelling in master planning on semiconductor supply chain performance.* Munich: Technische Universität München.

Yuan, J., & Ponsignon, T. (2014). Towards a semiconductor supply chain simulation library (SCSC-SIMLIB). *Proceedings of the 2014 Winter Simulation Conference* (pp. 2522-2532). Savannah, GA: IEEE.

Appendices

Appendix A: Stock target parameters in more detail

At Infineon, stock targets are set by the supply chain planner based on experience. Efforts to use an advanced tool that calculates stock targets based on forecast errors, lead time, supply variability, and service level were not successful. (Álvarez Luque, 2015)

In the current process, the parameters that need to be set are:

Product type

The supply chain planner can choose to set a target stock for a specific product or product group. Furthermore, the supply chain planner can choose different granularities, for instance material number, sales product or product family (plan position type). We explain these terms in more detail in Appendix F.

Manufacturing level

When defining the manufacturing level, the supply chain planner can choose where to store the targeted stock, for example at the die bank or at the DC.

Demand type

The supply chain planner can choose to exclusively consider orders, to use only forecasts, or to consider both orders and forecasts.

Unconstrained or constrained demand

Moreover, the supply chain planner can choose between using unconstrained or constrained demand. With unconstrained demand is meant the demand without constraints, such as capacity, so what could be sold to the customers when for example capacity is unlimited. This is the opposite of constrained demand.

Calculation period

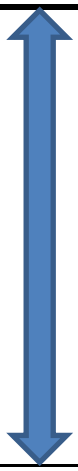
The supply chain planner also specifies the period for the average demand calculation.

Stock type

For the stock type parameter, Infineon defined three different stock types and because the system has to know what to plan first, a priority is given to each of these stock types. The

stock types and their priorities are shown in Table A-1. For each of these stock types, stock targets can be set.

Table A-1 Stock types and their priority

Stock type	Definition	Priority
Safety stock	Inventory that protects against stock-outs due to fluctuations in demand and supply.	 High
Ramp-up stock	Inventory that protects against an imminent increase in demand for new product ramp-ups.	
Min stock	This inventory can be used to (partly) fulfil a customer order. The min stock is replenished when sufficient capacity is available.	Low

Appendix B: Clarification of Overall Equipment Effectiveness terms

Machine state terms (Infineon Technologies, 1998)

Downtime

Downtime can either be scheduled, non-scheduled or unscheduled.

- Scheduled downtime is the scheduled time or proportion of the overall equipment time the equipment is not able to perform its intended function. This includes among others (preventive) maintenance, setups, and production tests.
- Non-scheduled downtime is the unscheduled time or proportion of the overall equipment time equipment is not able to perform its intended function due to unplanned down events, for example in case of technical failures, or due to a too low quality of input materials.
- Unscheduled downtime is the time or proportion of the overall equipment time the equipment is not utilised in a schedule to be utilised in production, for example during holidays, weekends, or when the equipment needs to be newly installed or relocated (Infineon Technologies, 2008).

Non-sales production

Non-sales production consists of time the equipment produces units for in-house use, mainly for research and development (R&D).

Sales production

The sales production is the time the equipment produces units for sale.

Standby time

The standby time is the remaining time and can be defined as the time the equipment is not used for unit processing or engineering, but is however available for production.

Appendix C: Operating curve theory

Operating curve theory is based on G/G/1/∞ queueing theory (Aurand & Miller, 1997). The G stands for a general arrival distribution (the first and second G can differ), the 1 stands for a one-machine station, and ∞ for unlimited capacity in the buffers (Hopp & Spearman, 2008).

The FF is usually calculated by the formula:

$$FF = \frac{CT}{RPT} \tag{16}$$

In queueing theory the RPT is often notated as E(S) or 1/μ, which stands for the mean service time.

The utilisation (UUm), or notated as ρ in queueing theory, is calculated using the productive time (PR) and the standby time (SB) by the formula:

$$UUm = \frac{PR}{PR + SB} \tag{17}$$

The utilisation and the FF are determined as an average per week. The operating point is subsequently determined in a graph with the FF on the y-axis and the utilisation on the x-axis.

The shape of the operating curve is dependent on the variability (α) for which a low value indicates a good line performance. Alpha represents the variability of all aspects regarding the production process and defines the form of the operating curve. Alpha (α) can be calculated using the FF and the utilisation (UUm):

$$\alpha = \frac{(FF - 1) * (1 - UUm)}{UUm} \tag{18}$$

This alpha can be seen as the overall queue variability. For a G/G/1/∞ queue this is equal to the squared coefficient of variation of the time between arrivals of a station plus the squared coefficient of variation of effective process time of a station (C_e^2) divided by two (Hopp & Spearman, 2008):

$$\alpha = \frac{C_a^2 + C_s^2}{2} \tag{19}$$

The curve can be drawn by calculating the FF by using the calculated alpha value (α) of a week and the utilisation (UUm), varying from 0 to 1:

$$FF = 1 + \alpha * \frac{UUm}{1 - UUm} \quad (20)$$

This formula can be derived from Kendall's notation for mean cycle times, in which the CT is most commonly notated as E(T) :

$$E(T) = E(S) * (1 + \frac{C_a^2 + C_e^2}{2} * \frac{\rho}{1 - \rho}) \quad (21)$$

Two examples of the operating curve are drawn in Figure A-1. Example 1: first, the Operating Point is drawn in the graph. This is done based on the FF and the utilisation that week, which for this example have the values of 3 and 0.6 respectively. To draw the curve with the help of Equation 20, we calculate the α value (Equation 18). In this case α has a value of 1.33. Example 2: again the Operating Point is drawn in the graph, based on a FF of 2.7 and a utilisation of 0.7. This leads to an alpha value of 0.73. The curve is then drawn with the help of Equation 20. As demonstrated by this curve, a lower FF can be achieved at higher utilisation when the variability (α) is low.

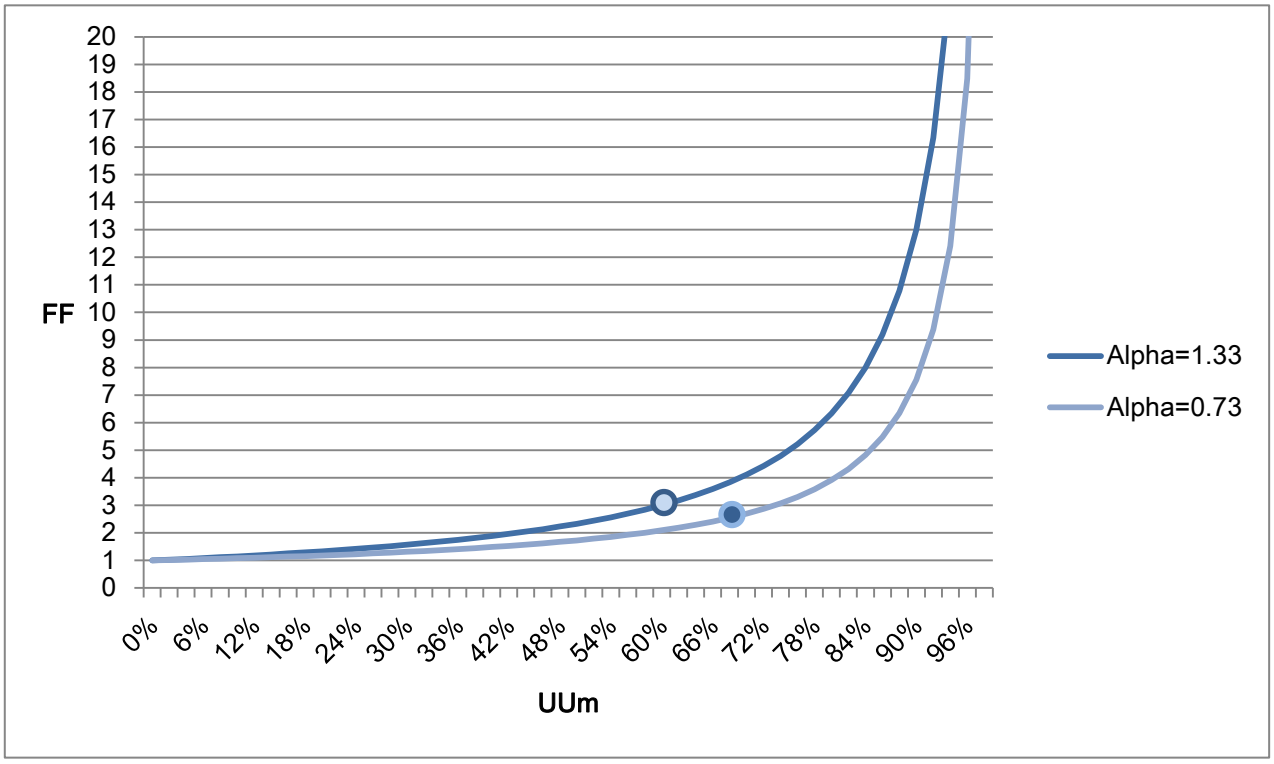


Figure A-0-1 Example of the Operating Curve

Appendix D: Plan function

In the planning part the release quantities per product into the FE (assuming infinite supply of raw materials) and from the die bank into the BE, are calculated backwards, considering WIP, actual and targeted inventory levels, orders, forecasts and unfulfilled production requests from previous weeks. The current version of the simulation model does not yet consider the Plan Load Limit, i.e. does not constrain the released quantities according to capacity restrictions.

The planning procedure is executed in the simulation model at the beginning of each week t with a planning horizon of 26 weeks. At this point in time, a snapshot of the current situation is taken. We explain below how the backward calculation of the release of lots works. An overview of the mathematical notations used can be found in Table A-2.

First we determine how many products should be released from the die bank into the BE based on backorders (unfulfilled requests from the DC) of the previous week $reqDC_{uf,\ t-1}$), WIP in the assembly (assyWIP), the test (testWIP), the BE yield (yBE) and the test yield (yTEST). The yield is the proportion of the number of units for sale to the total number of all units processed. Furthermore the release quantity is based on the demand that we expect at the beginning of period t during period t+τ (τ=0,1..25) ($D_{t,t+\tau}$) . τ is the planned CT in Assembly and Test and is given on a material number level (see Appendix F for more details for more details on these terms). Moreover, the targeted DC stock, which is based on the reach and the average demand pulled from the DC (dDC_{avg}), based on the expectation of the future demand, the reach set for the DC (RE_DC) in days, and the actual DC stock (stDC) contribute to the quality we want to release from the die bank into the BE. This has to be divided by the BE yield (yBE) to make sure the output corresponds the desired output.

The targeted quantity to be released for the die bank into the BE ($prodBE_{target}$) is calculated with the formula:

$$\begin{aligned} prodBE_{target,t} = \Bigg(& reqDC_{uf,t-1} + \sum_{\tau=0}^{BE\ CT} D_{t,t+\tau} \\ & - (assyWIP_t * yBE + testWIP_t * yTEST) \\ & + \frac{RE\_DC * dDC_{avg,t}}{7} - stDC_t \Bigg) / yBE \end{aligned} \tag{22}$$

where,

$$\sum_{\tau=0}^{BE\ CT} D_{t,t+\tau} = \sum_{\tau=0}^{BE\ CT-1} D_{t,t+\tau} + D_{t,t+BE\ CT} \tag{23}$$

Subsequently, the requested quantities required to be released into BE (reqBE) are determined. The requested quantities for BE manufacturing can get negative when, for example, the demand goes down or the yield is high. Therefore, the requested quantity gets a value of either zero or the minimum value of the maximum quantity that can be released from the die bank (maxRelDB), based on the actual die bank stock (stDB), and the targeted BE quantity multiplied by the number of chips to be produced in case of multichips (mc_{qty}). Multichips are assembled in the BE and consist, as the name already implies, of multiple chips. Thus, one multichip end-product consists of multiple semi-finished products. In case the end-product is not a multichip the value of mc_{qty} is one.

$$reqBE_t = max\left\{0, min\{maxRelDB_t, prodBE_{target,t}\& * mc_{qty}\}\right\} \quad (24)$$

where,

$$maxRelDB_t = stDB_t / mc_{qty} \quad (25)$$

It is not always possible to release all these requested products from the die bank, so there are also unfulfilled die bank requests ($reqDB_{uf}$) for this week (t); there is a discrepancy in what we want to release and what we can release. The released requests are also called to be fulfilled ($reqDB_f$). To back-up the unfulfilled requests, a targeted FE quantity to be produced should be determined. To determine this, first the demand expected after the FE CT period ($D_{t,t+FE\ CT}$) is determined by the formula:

$$D_{t,t+FE\ CT} = reqDB_{f,t} + reqDB_{uf,t} \quad (26)$$

After that, the targeted FE quantity to be produced for the die bank ($prodFE_{target}$) is calculated. The targeted products to be produced that week is dependent on the unfulfilled requests of the previous weeks, the demand that we expect at the beginning of period t during period t+τ (τ=0,1..25) ($D_{t,t+\tau}$). This τ depends on the planned FE CT per FE facility per product, WIP in the assembly (fabWIP). Furthermore, the targeted products to be produced that week is dependent on the test (sortWIP, the number of chips that need to be produced in case of multichips (mc_{qty}), the targeted die bank stock which is based on the reach, and the average demand pulled from the die bank (dDB_{avg}) based on the expectation of the future demand, the reach set for the die bank (RE_DB), and the actual die bank stock (stDB).

$$prodFE_{target,t} = (reqDB_{uf,t-1} + \sum_{\tau=0}^{FE\ CT} D_{t,t+\tau} * mc_{qty} - (fabWIP_t + sortWIP_t) + \frac{RE\_DB * dDB_{avg,t} * mc_{qty}}{7} - stDB_t) \quad (27)$$

where,

$$\sum_{\tau=0}^{FE\ CT} D_{t,t+\tau} = \sum_{\tau=0}^{FE\ CT-1} D_{t,t+\tau} + D_{t,t+FE\ CT} \quad (28)$$

Subsequently, the requested quantities to be released into the FE (reqFE) are determined:

$$reqFE_{f,t} = max \begin{cases} 0 \\ prodFE_{target,t} \end{cases} \quad (29)$$

Afterwards, the lots that need to be started in the FE (lsFE) are calculated by taking the yield (yFE), chips per wafer (cpw) and the lot size in the FE (lsizeFE) into account:

$$lsFE_t = \left\lceil \frac{ReqFE_{f,t}}{yFE * cpw * lsizeFE} \right\rceil \quad (30)$$

Finally, the lots that need to be started in the BE (lsBE) are calculated by dividing the fulfilled die bank requests by the lot size in the BE (lsizeBE):

$$lsBE_t = \left\lceil \frac{ReqDB_{f,t}}{lsizeBE} \right\rceil \quad (31)$$

The lot sizes used in the existing version of the simulation model are equal for each product, period and each facility. There is only differentiated between BE and FE lot sizes.

Table A-2 Mathematical notation backwards calculation lot starts

Notation	Explanation
assyWIP	WIP in 'assembly'
cpw	Number of chips per wafer
$dBECT_{after}$	Demand after the BE CT: the production units demanded that have to be started to be produced in time in the BE
$dBECT_{during}$	Demand during the BE CT: the demand that should be covered by releases into the BE in one of the previous weeks
dDB_{avg}	Average demand pulled from the die bank
dDC_{avg}	Average demand pulled from the DC
$dFECT_{after}$	Demand after the FE CT: the production units demanded that have to be started to be produced in time in the FE
$dFECT_{during}$	Demand during the FE CT: the demand that should be covered by releases into the FE in one of the previous weeks
fabWIP	WIP in 'fab'
lsBE	Number of lots to be started in the BE
lsFE	Number of lots to be started in the FE
lsizeBE	Lot size in the BE
lsizeFE	Lot size in the FE
maxRelDB	Maximum quantity that can be released from the die bank
mc_{qty}	Number of chips to be produced in case of multichips
$prodBE_{target}$	Targeted BE quantity to be produced in the BE for the DC
$prodFE_{target}$	Targeted FE quantity to be produced for the die bank
RE_DB	Reach for the die bank
rDC	Reach for the DC
reqBE	Requested quantities to be fulfilled by production in the BE
$reqDB_{f}$	Fulfilled die bank requests
$reqDB_{uf}$	Unfulfilled die bank requests
$reqDC_{uf}$	Unfulfilled DC requests (backorders)
reqFE	Requested quantities to be fulfilled by production in the FE
sortWIP	WIP in 'sort'
stDB	Actual die bank stock
stDB	Actual die bank stock
stDC	Actual DC stock
testWIP	WIP in 'test'
yBE	BE yield
yFE	FE yield

Appendix E: Non-bottleneck delay determination (existing model)

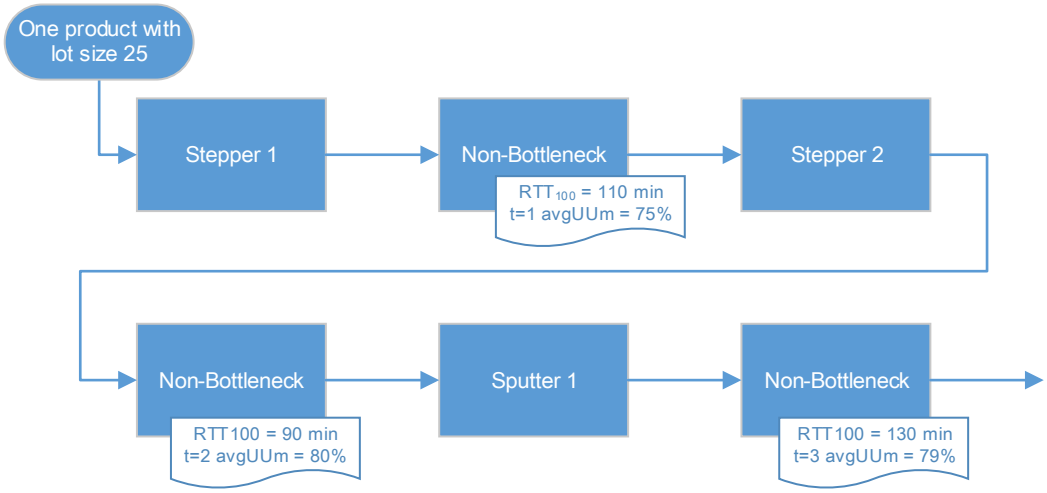


Figure A-0-2 Process of product X for the delay of non-bottleneck determination example

For the first non-bottleneck step we calculated a delay of 93.5 minutes. Now we have to determine the second and third non-bottleneck step delays (as shown in Figure A-2). The alpha remains 0.38.

When a lot arrives at the second non-bottleneck step (t=2), we calculate the FF:

$$FF = 1 + \alpha * \frac{UUm}{1 - UUm} = 1 + 0.38 * \frac{0.80}{1 - 0.80} = 2.5$$

Before we can calculate the non-bottleneck delay for this second step we also have to determine the raw processing time, by using the given raw tool time for that step. The scaleRPT is again assumed to be 1.6, which is also the assumption in the simulation model.

$$RPT = scaleRPT * \frac{RTT_{100} * lot\ size}{100} = 1.6 * \frac{90 * 25}{100} = 36\ minutes$$

Now we calculate the delay of the second non-bottleneck step:

$$DelayNB = RPT * FF = 36 * 2.5 = 90\ minutes$$

Note that the sputter utilisation is not taken into account for the third non-bottleneck step delay (t=3), however, the average UUm at t=3 is 0.79. Therefore, we have to calculate the FF again:

$$FF = 1 + \alpha * \frac{UUm}{1 - UUm} = 1 + 0.38 * \frac{0.79}{1 - 0.79} = 2.41$$

We determine the raw processing time, by using the given raw tool time for that step. The scaleRPT is still assumed to be 1.6.

$$RPT = scaleRPT * \frac{RTT_{100} * lot\ size}{100} = 1.6 * \frac{130 * 25}{100} = 52\ minutes$$

Now we can calculate the delay of the third non-bottleneck step:

$$DelayNB = RPT * FF = 52 * 2.41 = 125.32\ minutes$$

Appendix F: Product master data

The product master data includes the material number, the sales product, the construction number, the number of manufacturing routes, the product family, the position in the facility chain, the facility type, the facility code, the manufacturing level, the location abbreviation, the planned CT, the multi-chip quantity, the technology class, the technology, the process class, the process group, the process line, the basic type identifier of the items in a specific facility, the basic type identifier for the DC, the manufacturing route number, the number of lithography layers (stepper visits), the package class, the package group, the package, the division, the business line and the product line. These terms are clarified in Table A-3.

Table A-3 Clarification of product master data

Term	Abbreviation used in Infineon's databases	Clarification
Material Number	MA	The product number used within Infineon. Every MA can be linked to one SP.
Sales Product	SP	The product number communicated to the customer. One SP can be linked to multiple MAs.
Construction Number	BNR	The construction number changes every manufacturing level.
Manufacturing route	MR	In case of a multichip, the different chips might have to follow different routes before they are assembled.
Product family	PPOS type	Distinguishes between chips, bare dies, multichips and devices. The chips and bare dies are processed only in the FE, the multichips and devices in both the FE and BE.
Position in the facility chain	Pos IF	The process steps from DC counted backwards.
Facility type	Facility type	Distinguishes between facilities (=F) and warehouses (=L).
Facility code	Facility	The facility names in code: Facilities - Regensburg = _1702, Villach = _1502, Kulim = _WFKUL. Die banks - Regensburg = _LGDPR, in Villach = _LAGERV, in Kulim = _LGK, and many others (not relevant). DCs/Hubs - Europe = DCE, China = DCC, Asia = DCA, USA = DCU.

Manufacturing level	ML	Refers to the steps in the supply chain: fab = FAB, sort = SORT, die bank = DIEBANK, assembly = ASSY, test = TEST and DC = VKL.
Location abbreviation	Loc key	The abbreviated facility names: Facilities & die banks - Regensburg = RBG, Villach = VIH, Kulim = KLM. DCs - Europe = DCE, Asia = DCA, USA = DCU, China (local hub only) = DCC
Planned cycle time	Planned CT	The planned cycle time in days for a specific product for a specific manufacturing level.
Multi-chip quantity	FE qty	The quantity of the corresponding basic type (IF) in the final product.
Technology class	TC	Technology class
Technology	Technology	Distinguishes, among others, between the power and CMOS technology classes.
Process class	Process class	Letter-number combinations that stand for the aggregation of process groups.
Process group	Process group	Letter-number combinations that stand for the aggregation of similar process lines.
Process line	Process line	Name within Infineon's global data system for all unit process steps to be performed in order to manufacture a product.
Basic type identifier of the items in a specific facility	Basic type (IF)	A logistical identifier for the items in a specific facility.
Basic type identifier for the DC	Basic type (DCBNR)	A logistical identifier for delivery to a DC.
Manufacturing route number	Workroute	Letter-number combinations that stand for a specific route in that facility.
Number of lithography equipment visits	Litho steps	The number of lithography steps needed for that product in that facility.
Package class	Package class	Letter-number combinations that stand for the aggregation of package groups.
Package group	Package group	Letter-number combinations that stand for the aggregation of similar process packages.
Package	Package	The letter-number combinations that stand for chips that are assembled the same way.
Division	Division	Defines for which division of Infineon the product is produced. Distinguishes between Automotive (ATV), Power Management

		& Multimarket (PMM), Industrial Power Control (IPC), Chip Card & Security (CCS) or Other Operating Segment (OOS).
Business line	BL	The number identification of the business line. A business line consists of one or multiple product lines.
Product line	PL	The number identification of the product line. One product line is linked to a business line.

Appendix G: Route master data

To explain how the routing in the simulation model is determined, an example is given. Part of the master data of a fictional product with MA001 is shown in Table A-4.

Table A-4 Part of master data of a product with material number 001

MA	BNR	PPOS type	Pos IF	ML	Loc Key	CT (plan)	Technology	Process Group	Basic Type (DCBNR)	Workroute
MA001	00000001	Chip	14	FAB	VIH		PT	ST1D_8	A0001A	AA-000AA1
MA001	00000002	Chip	10	FAB	RBG		WAFERFINISH	SE3EIR	A0001A	SE-000SE1
MA001	00000003	Chip	6	FAB	VIH	21	WAFERFINISH	AN3ADR	A0001A	AA-000AA2
MA001	00000004	Chip	5	SORT	VIH	7			A0001A	
MA001	00000004	Chip	4	DIEBANK	VIH				A0001A	
MA001	00000005	Chip	1	VKL	DCE				A0001A	

For this product, wafers are first processed in Villach, second in Regensburg, and third, again in Villach. After that, they stay in Villach for the 'sort' step and are stored at the die bank at the same location. From the die bank they are transported to the Europe DC. This product is not processed in the BE facilities, because the PPOS type is 'chip'. In these facilities, the wafers each follow a different route, which is indicated with a special letter-number combination. This is visualised in Figure A-3.

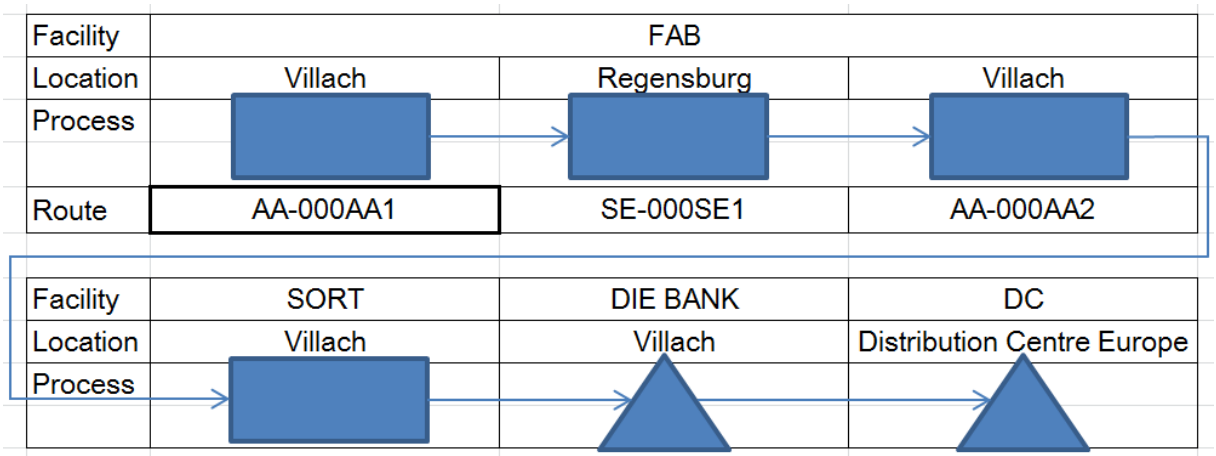


Figure A-0-3 Visualisation of the process of a product with material number 001

The letter-number combination for the route indication corresponds with a more detailed route-information the route table. Here is indicated which equipment the wafers visit in which sequence. For MA001 the route-information for the first process in Villach, with letter-number combination AA-000AA1, can be found in Table A-5. As can be read in this table, also the Raw Tool Times (RTT) of this product on each of the equipment are stated, as well as the facility code. Which is in the case of Villach 1502.

Table A-5 Route-information route PD-4065V1 (Villach)

ROUTE	BOTTLENECK_AGGREGATED	RTT (min/100 wafer)	FACILITY
AA-000AA1	NONBN	1196.2	1502
AA-000AA1	LITHO/STEPPER	80	1502
AA-000AA1	NONBN	3694.2	1502
AA-000AA1	LITHO/STEPPER	106.6	1502
AA-000AA1	NONBN	2576.8	1502
AA-000AA1	SPUTTERN	174	1502
AA-000AA1	NONBN	121	1502
AA-000AA1	LITHO/STEPPER	151	1502
AA-000AA1	NONBN	1614.2	1502
AA-000AA1	LITHO/STEPPER	178	1502
AA-000AA1	NONBN	6459.3	1502

The specified route on the equipment in Villach and its corresponding Raw Tool Time is visualised in Figure A-4.

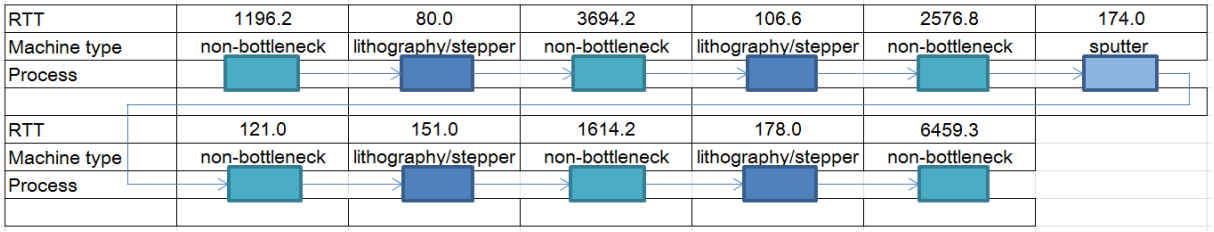


Figure A-0-4 Visualisation of route PD-4065V1 in Villach

This is how the routing for every product is determined by the simulation model.

Appendix H: Key performance indicators for validation

Data	Definition
WSPW	Number of wafers started
LSPW	Number of layers the started wafers need to get (number of lithography visits)
WOPW	Number of wafers leaving the fab
LOPW	Number of layers (on wafers) leaving the fab
Utilisation	Loading of stepper and lithography equipment; the share of the productive time of the uptime
WIP	The number of product units that entered the wafer fab, but have not left the fab yet
CT	The time spent by a product unit in the wafer fab, from the release of the wafer into the fab till leaving the wafer fab manufacturing level
FF	The flow factor measured at the end of the process, so from entering the first equipment the wafer fab until leaving the last equipment in the wafer fab. The CT and RPT are used to calculate this factor

Appendix I: Implementation improvements

Key performance indicators calculation alignment

First three scenarios are set-up to determine if there is any difference in the calculation of the WSPW, WOPW, LSPW, LOPW, and the CT in different cases. Secondly we show the experts these KPIs measurements we believe are correct. This is both done in a structured and visual way so that it is easily understandable for the database experts.

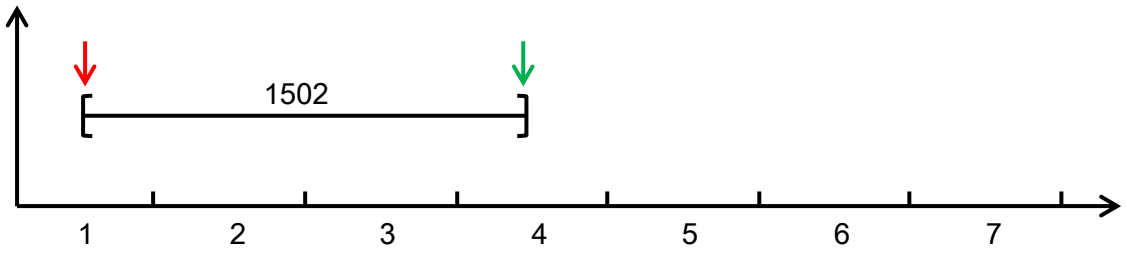
One lot with lot size 25 enters the system. The route of this lot is dependent on the scenario:

- Scenario 1: one facility visit
  - Villach
  - Villach – CT = 3 weeks – 7 lithography layers
- Scenario 2: two facility visits
  - Villach → Regensburg
  - Villach – CT = 3 weeks – 7 lithography layers
  - Regensburg – CT = 2 weeks – 1 lithography layer
- Scenario 3: three facility visits
  - Villach (1) → Regensburg → Villach (2)
  - Villach (1) – CT = 3 weeks – 7 lithography layers
  - Regensburg – CT = 2 weeks – 1 lithography layer
  - Villach (2) – CT = 1 week – 0 lithography layer

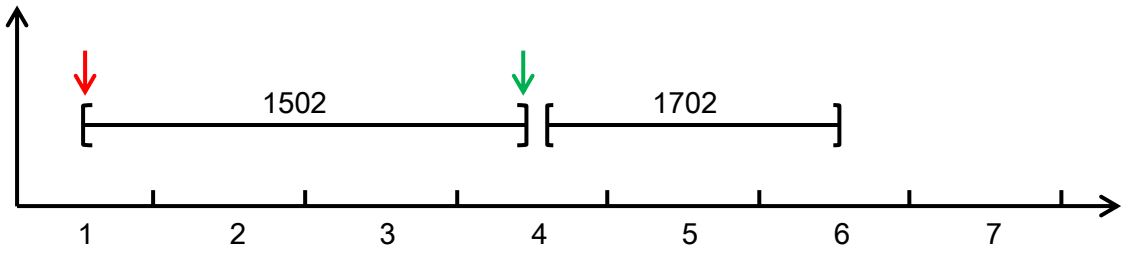
The actual KPIs appear to be measured in the following way:

- WSPW (red) & WOPW (green)

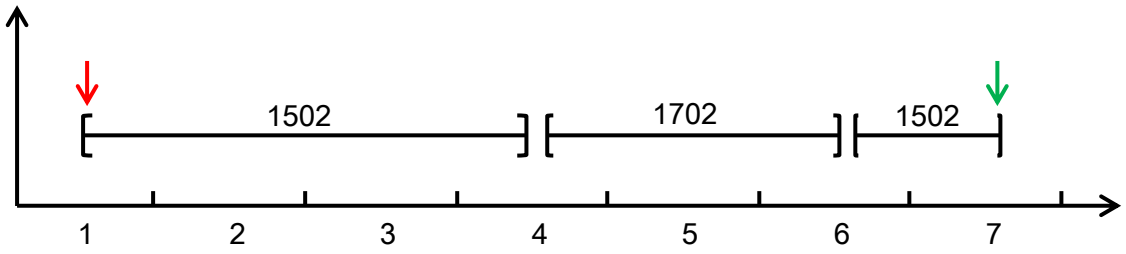
  Scenario 1:



  Scenario 2:



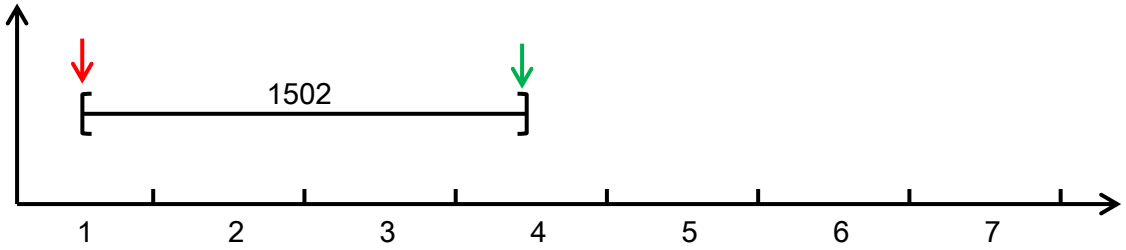
Scenario 3:



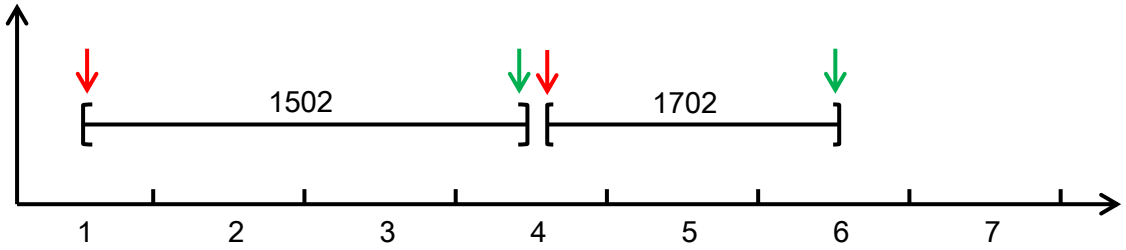
The WSPW is measured each time a lot enters the first wafer fab facility and is written to that facility. The WOPW is measured when a lot leaves the first facility for the last time in the route and written to that first facility.

- LSPW (red) & LOPW (green)

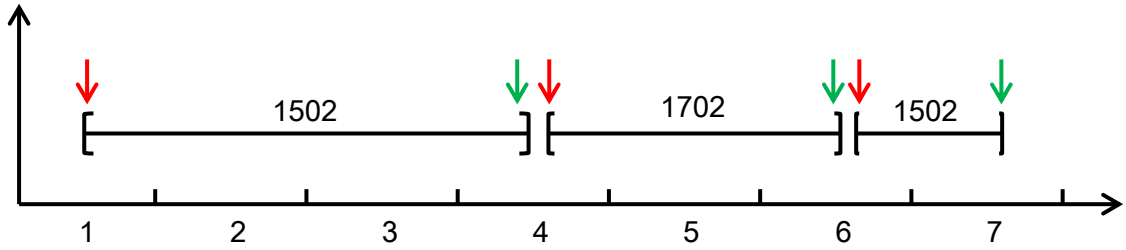
  Scenario 1:



Scenario 2:



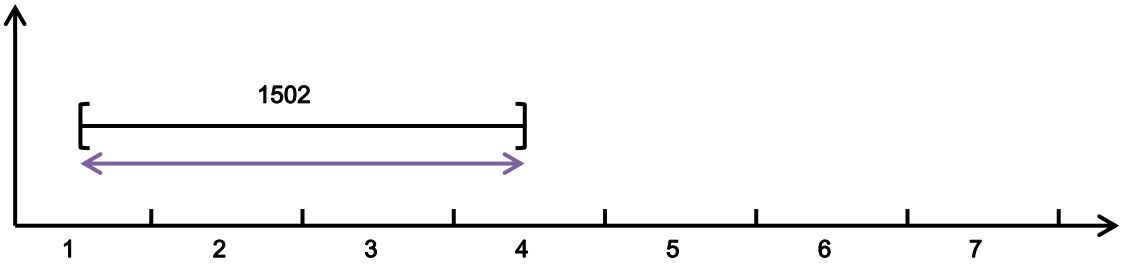
Scenario 3:



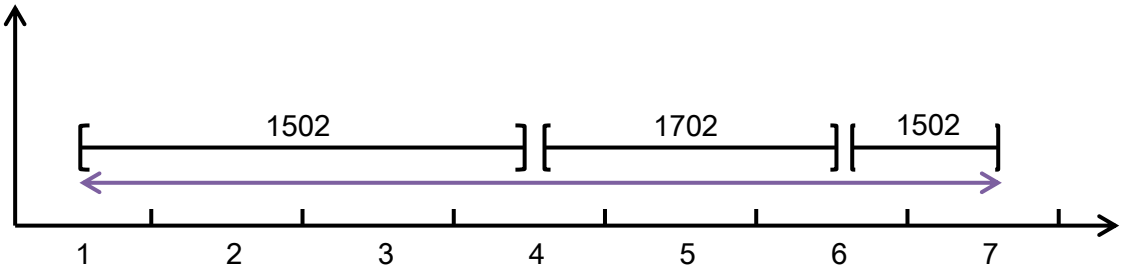
The LSPW is measured each time a lot enters a wafer fab facility and written to that facility. The WOPW is measured every time a lot leaves a facility and written to that facility as well.

- CT (purple)

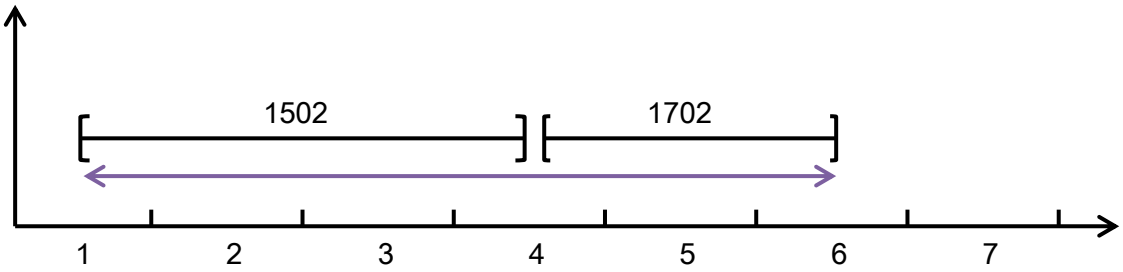
  Scenario 1:



  Scenario 2:



  Scenario 3:



The CT is measured from entering the first facility until leaving the last facility. The CT is written to the last facility visited.

The FF and utilisation are calculated according to the corresponding formulas, Equation 16 and Equation 17 of Appendix C respectively. The WIP is measured by adding the number of wafers to that variable when they enter a facility and subtract the number of wafers that leave a facility. This is measured for each facility, independent of the route.

Flexible bottleneck implementation

Two main adjustments have been necessary to implement the flexible bottlenecks.

On initialisation the bottlenecks are read into the simulation model by a function called “setBottlenecksFE”. The programming language used by AnyLogic is Java.

```
//for all facilities
for(int j=0; j<facilitiesFE.size();j++){
   //store all facility names in a collection
```

```
String facilityFE = facilitiesFE.get(j).siteName;

//initialize "count"
int count = 0;

//for all indicated bottlenecks.
for(int i = 0; i<(inputsExternal.bottlenecks_FE_DB.size());i++){

    //when the facility equals the facility in the access input table
    if(inputsExternal.bottlenecks_FE_DB.get(i).facility.equals(facilityFE)){
       //store the name of all the bottlenecks
      String bottleneckName =
      inputsExternal.bottlenecks_FE_DB.get(i).bottleneck;

     //the bottleneck name is equal to the bottleneck in the access input table
     if(inputsExternal.bottlenecks_FE_DB.get(i).bottleneck.equals(bottleneckNam
     e)){

       //add the bottleneck to the bottlenecks in productionUnitFE and add to
       the bottlenecksFE collection
       count += 1;

       //bottleneck agent is 1 by default
       if (count > 1){
         //add bottlenecks to the bottleneck work center for each facility
         facilitiesFE.get(j).productionUnitFE.add_bottlenecks();
       }

       //fill a collection to be able to indicate the name of the bottleneck
       related to the index the bottleneck has
       facilitiesFE.get(j).productionUnitFE.bottleneckNamesIndices.put(bottlene
       ckName, count-1);

       //set a parameter to be able to indicate the name of the bottleneck
       facilitiesFE.get(j).productionUnitFE.bottlenecks.get(facilitiesFE.get(j)
       .productionUnitFE.bottlenecks.size()-1).nameWorkCenter = bottleneckName;

       //set the Raw Tool Time factor
       facilitiesFE.get(j).productionUnitFE.bottlenecks.get(facilitiesFE.get(j)
       .productionUnitFE.bottlenecks.size()-1).factorRTT =
       inputsExternal.bottlenecks_FE_DB.get(i).factor;

       //add bottlenecks to facilitiesFE collection on main layer
       int max_bn_fac =
       (facilitiesFE.get(j).productionUnitFE.bottlenecks.size()-1);
       bottlenecksFE.add(facilitiesFE.get(j).productionUnitFE.bottlenecks.get(m
       ax_bn_fac));

       //set number of tools, availability for each bottleneck of each facility
       int tools =
       inputsExternal.tool_qty_FE.get(facilityFE).get(bottleneckName).get(1);
       double availabilityBottleneck =
       inputsExternal.availability_FE.get(facilityFE).get(bottleneckName).get(1
       );

       facilitiesFE.get(j).productionUnitFE.bottlenecks.get(facilitiesFE.get(j)
       .productionUnitFE.bottlenecks.size()-1).numberTools = tools;
```

```
            facilitiesFE.get(j).productionUnitFE.bottlenecks.get(facilitiesFE.get(j)
            .productionUnitFE.bottlenecks.size()-
            1).resourcePool.set_capacity(tools);

            facilitiesFE.get(j).productionUnitFE.bottlenecks.get(facilitiesFE.get(j)
            .productionUnitFE.bottlenecks.size()-1).numberToolsVar = tools;

            facilitiesFE.get(j).productionUnitFE.bottlenecks.get(facilitiesFE.get(j)
            .productionUnitFE.bottlenecks.size()-1).availability =
            availabilityBottleneck;

            //read the real utilisation of the bottlenecks from the input data to
            compare the simulated utilisation in a graph per bottleneck
            double utilisationCerberus =
            inputsExternal.uum_FE.get(facilityFE).get(bottleneckName).get(1);

            facilitiesFE.get(j).productionUnitFE.bottlenecks.get(facilitiesFE.get(j)
            .productionUnitFE.bottlenecks.size()-1).utilCerberus =
            utilisationCerberus;

            }
        }
    }
}

//make a reference to the routes for each bottleneck
for(int i = 0;i<bottlenecksFE.size();i++){
       bottlenecksFE.get(i).MasterData_Ref.refToRoutes = routes;
}
```

The parameters set need to be updated when the simulation model is running. This is done in the function “adjustParameters”. This function is called at the beginning of every simulated week.

```
//initialize "count"
int count = 0;

//for all facilities
for(int i=0;i<facilitiesFE.size();i++){

   //get facility name
   String facilityFE = facilitiesFE.get(i).siteName;

   //flexible bottlenecks
   count = 0;

   //for all bottlenecks
   for(int j = 0; j<(inputsExternal.bottlenecks_FE_DB.size());j++){
      //when the facility equals the facility in the access input table
      if(inputsExternal.bottlenecks_FE_DB.get(j).facility.equals(facilityFE)){
      //store the name of the bottleneck
      String bottleneckName = inputsExternal.bottlenecks_FE_DB.get(j).bottleneck;

         //when the facility equals the facility in the access input table
```

```
            if(inputsExternal.bottlenecks_FE_DB.get(j).bottleneck.equals(bottleneckNa
            me)){

                //update the tools and availability of the bottlenecks
                int toolsBottleneck =
                inputsExternal.tool_qty_FE.get(facilityFE).get(bottleneckName).get(we
                ekAfterWarmUp);
                double availabilityBottleneck =
                inputsExternal.availability_FE.get(facilityFE).get(bottleneckName).ge
                t(weekAfterWarmUp);


                facilitiesFE.get(i).productionUnitFE.bottlenecks.get(count).numberToo
                ls = toolsBottleneck;

                facilitiesFE.get(i).productionUnitFE.bottlenecks.get(count).resourceP
                ool.set_capacity(toolsBottleneck);

                facilitiesFE.get(i).productionUnitFE.bottlenecks.get(count).numberToo
                lsVar = toolsBottleneck;

                facilitiesFE.get(i).productionUnitFE.bottlenecks.get(count).availabil
                ity = availabilityBottleneck;

                //update "count"
                count += 1;

            }
        }
    }
}
```

Delay determination approaches implementation

Less aggregated alpha

Alpha calculation by creating a Microsoft Access query:

The query contains the facility, the week, the average utilisation of the facility-specific bottlenecks, the process group, the FF, and the calculated alpha based on this data. This is based on Equation 18 in Appendix C. The following expression is implemented in the query:

Alpha: ((([FF]-1)*(1-[AvgOfUUM_AVG]))/[AvgOfUUM_AVG])

The function to load this alpha value in the simulation model is called “getAlpha”:

```
//initialize alpha value and get product
double alpha = 0.5;
int pos_if = lot.facilityCount-1;
int mr = lot.mr;
String ma = lot.ma;
String facilityName =
readFEInputs.masterData.masterDataProducts.get(ma).facility.get(mr).get(pos_if);
```

```
//if a less aggregated alpha value is available in the input file
if
(readFEInputs.inputsExternal.alpha_FE.get(lot.firstFab).containsKey(lot.firstProce
ssGroup)&&readFEInputs.inputsExternal.alpha_FE.get(lot.firstFab).get(lot.firstProc
essGroup).containsKey(readFEInputs.weekAfterWarmUp)){

    //alpha takes that value
    alpha =
    readFEInputs.inputsExternal.alpha_FE.get(lot.firstFab).get(lot.firstProcessGro
    up).get(readFEInputs.weekAfterWarmUp);
}
//if there is no less aggregated alpha value available in the input file
else{
      //for all facilities
      for(int i=0;i<readFEInputs.facilitiesFE.size();i++){

            //get facility name
            String facilityFE = readFEInputs.facilitiesFE.get(i).siteName;

            //when the facility equals the facility in the access input table
            if (facilityFE.equals(lot.firstFab)) {
                  alpha = readFEInputs.facilitiesFE.get(i).alpha_PRISM;
                  break;
            }
      }
}
return alpha;
```

Use all bottleneck utilisations to determine non-bottleneck delay

For all bottlenecks the utilisation is stored in a variable called “uUm_now”, a dashboard called “statUUm_now” collects these statistics. The “uUm_now” is calculated based on the productive time and the standby time of the equipment (Equation 2, Appendix C).

The function is triggered by an event that makes sure it is updated every four hours:

```
//initialize “aveUtilTotalBottlenecks”
double aveUtilTotalBottlenecks = 0;

//for all bottlenecks
for (int i = 0; i<bottlenecks.size(); i++){
      //get the mean uptime utilisation for manufacturing and add it to
      “aveUtilTotalBottlenecks”
      aveUtilTotalBottlenecks += bottlenecks.get(i).statUUm_now.mean();
      //reset the statistic for that bottleneck
      bottlenecks.get(i).statUUm_now.reset();
      //add value including time to the statistics
      bottlenecks.get(i).statUUm_now.add(bottlenecks.get(i).uUm_now, time());
}

//average utilisation is the sum of all uptime utilisations divided by the number
of bottlenecks
double aveUtil = aveUtilTotalBottlenecks/bottlenecks.size();

//add this value to a collection with historical delay data
utilForDelayHistory.add(aveUtil);
```

Implementation of lot size determination improvement

The number of lots are determined by creating a Microsoft Access query:

The query contains the week, the facility, the basic type, the material number, the manufacturing route, the number of started wafers, the number of lots with lot size 50, the number of lots with lot size 25, the number of lots with another lot size (the remainder wafers that did not fit in a lot of 25 or 50), the specified lot size for Villach based on historical data, the standard lot size for Villach if there is no specified lot size available, the started lots for Villach, the overall started lots, and the average lot size. The last two values are used in the simulation model.

The following expressions are implemented in the query:

Lots_Lotsize_50: IIf([FACILITY_NAME]="1702";Int(([STARTED_WAFERS]*(2/3))/50);0)

Lots_Lotsize_25:
IIf([FACILITY_NAME]="WFKUL";Int([STARTED_WAFERS]/25);IIf([FACILITY_NAME]="1702";Int(([STARTED_WAFERS]-([Lots_Lotsize_50]*50))/25);0))

Lots_Lotsize_Other:
IIf([FACILITY_NAME]="WFKUL" Or [FACILITY_NAME]="1702";IIf([STARTED_WAFERS]-([Lots_Lotsize_50]*50+[Lots_Lotsize_25]*25)>0;1;0);0)

Lotsize_Vil:
IIf([FACILITY_NAME]="1502";IIf(IsNull([Lotsize_SPEC]);[Standard_Lotsize_Vil];[Lotsize_SPEC]);0)

Standard_Lotsize_Vil: IIf([Facility_Name]="1502";50;0)

STARTED_LOTS_INT_Vil:
IIf([Facility_Name]="1502";-Int(-([STARTED_WAFERS]/[Lotsize_Vil]));0)

STARTED_LOTS_INT:
[Lots_Lotsize_50]+[Lots_Lotsize_25]+[Lots_Lotsize_Other]+[STARTED_LOTS_INT_Vil]

AVE_LOTSIZE: [started_wafers]/[started_lots_int]

Facility “1502” is Villach, facility “1702” is Regensburg and facility “WFKUL” is Kulim.

Appendix J: Warm-up period (first warm-up phase)

As we start with an empty system we want to fill the model in such a way that it reaches a steady-state before starting the experiments. By this end, the initialisation bias will be reduced. The time the simulation models needs to reach a steady state we refer to as the warm-up period. In the first phase only the input data of the first week is sent into the system. After this period, the experiment starts with the historical data of the other weeks.

For the determination of the first warm-up phase we follow the steps of Welch's graphical method (Law, 2007).

Since our purpose with this warm-up phase is to fill the model, we run a simulation in which the output is measured in WOPW per facility with a length (j) of 150 and make 10 independent replications (r):

$$WOPW_{jr}\ (j = 1 \dots 150, r = 1 \dots 10) \tag{32}$$

After that, we calculate the mean of the ith observation over 10 runs:

$$\overline{WOPW_j} = \frac{1}{10} \sum_{r=1}^{10} WOPW_{ij} \tag{33}$$

Next, we take average over a window (w) of 2 to smooth out high-frequency oscillations:

$$\overline{WOPW_j}(2) = \frac{1}{4+1} \sum_{s=-2}^{2} \overline{WOPW}_{j+s} \tag{34}$$

If i ≤ 2 we use w = i-1.

Thereafter, we plot the moving averages and choose observation h beyond which the output seems to be stable. We plot the moving averages for both the simulation existing and the updated simulation model, as displayed in Figure A-5 and Figure A-6 respectively. The Figures show that the output stabilises after 52 weeks or less. Therefore, we use a first warm-up phase of 52 weeks for our experiments.

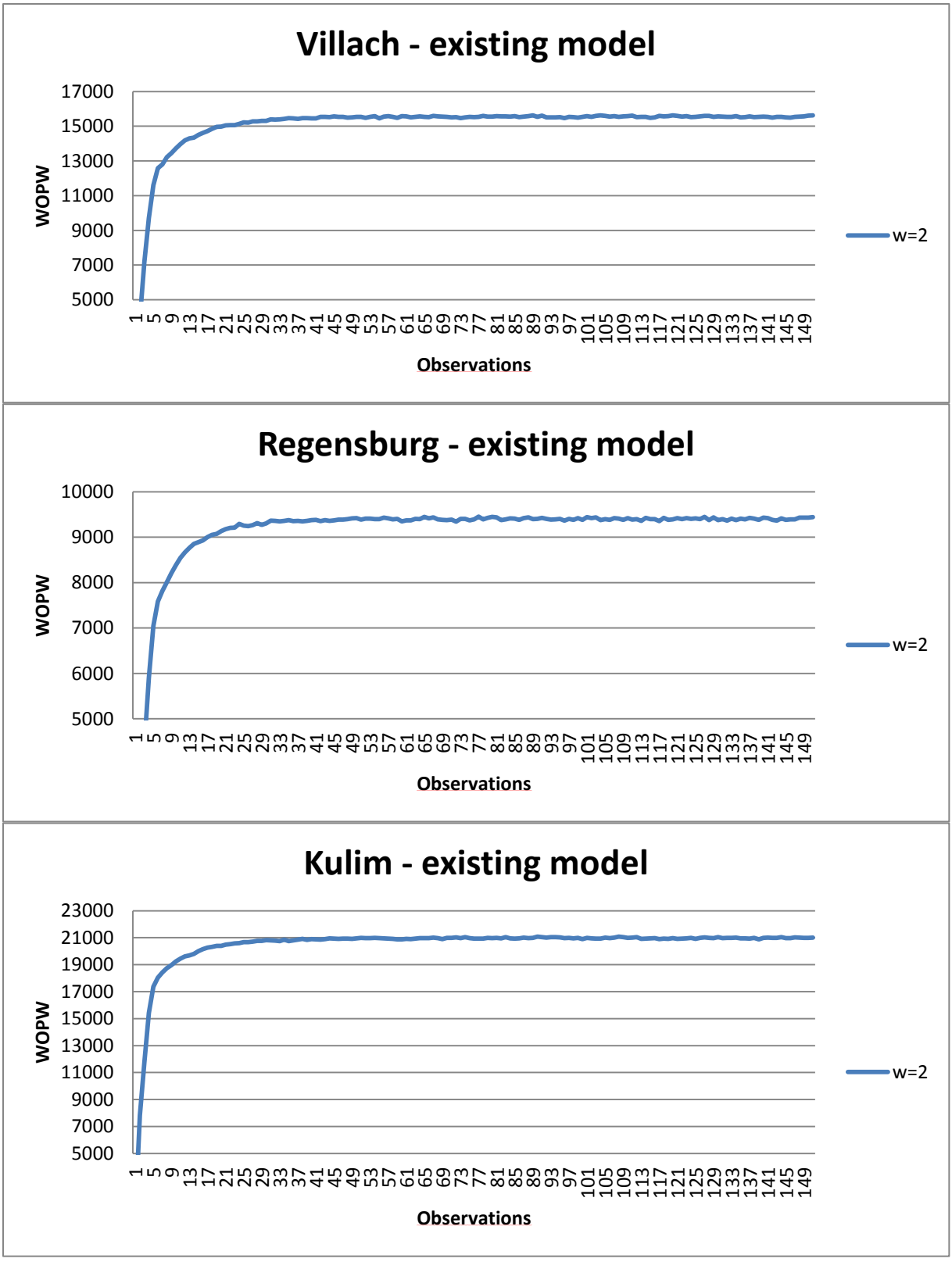


Figure A-0-5 Welch's graphical method for the existing simulation model

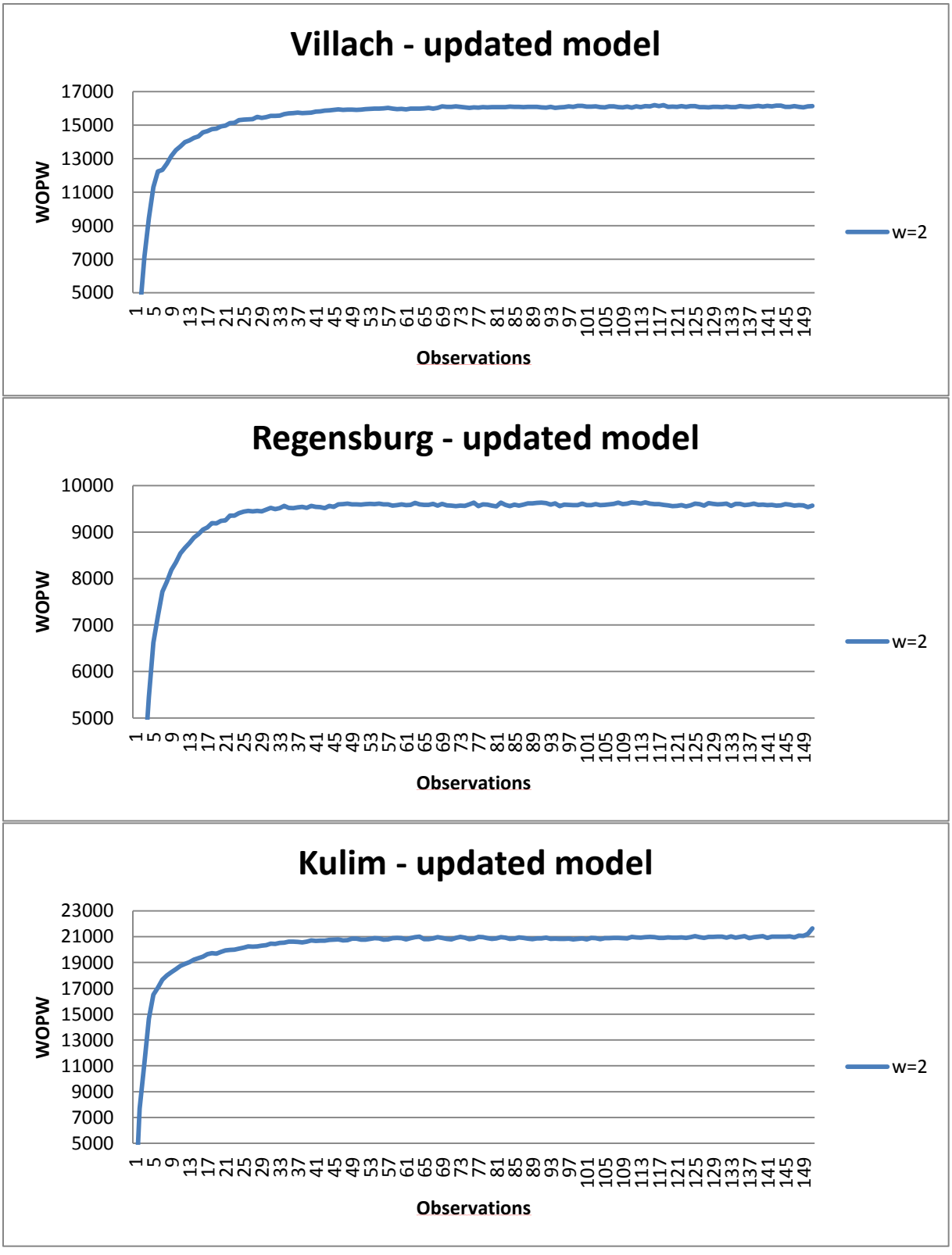


Figure A-0-6 Welch's graphical method for the updated simulation model

Appendix K: Replications

To gain reliable results for our experiments, we need to replicate the simulation a certain number of times to ensure the results lie in the desired confidence interval with a small relative error. Our goal is to achieve a confidence interval of at least 95% (α = 0.05) and a relative error of less than 5% (γ = 0.05) for the output of each facility. We determine this for each facility for both the existing and updated simulation model.

We run the model with a warm-up period of 52 weeks and continue the run for 78 weeks (i = 1…78) with historical input data. We make, again, 10 independent replications (r = 1…10) of this run and analyse the output of the WOPW.

We have to seek for the minimal n* for which the corrected target value ≤ γ/(1+γ) (Law, 2007) by Equation 35. To do this, we need the student-t distribution ($t_{i-1,1-\alpha/2}$), the average over the replications ($\overline{X}_r$), and the variance (S_r^2) over the replications.

$$n* = \min\left\{ t \geq n: \frac{t_{i-1,0.975}\sqrt{\frac{S_n^2}{i}}}{|\bar{X}_n|} \leq \frac{0.05}{1 + 0.05} \right\} \quad (35)$$

We determine n* by using a sequential procedure. This means that we calculate the values for every n, until we find a value which is smaller than the corrected target value which leads to our desired relative error.

Table A-6 summarises the results we obtained by this method. To ensure that we reach the relative error requirement and the desired confidence interval, we have to take the maximum number of replications found. According to our results in Table A-6, we need to make 2 replications to obtain reliable results.

Table A-6 Required number of replications for each facility for α = 0.05 and γ=0.05

	Villach	Regensburg	Kulim
Existing model	1	1	2
Updated model	1	1	2

As this number is low, we also decided to check the number of replications if we want to obtain a relative error of at most 0.5% (γ = 0.005) and a confidence interval of at least 99.5%

(α = 0.005). These results can be found in Table A-7. This table shows that we need to make at least 9 replications to reach this new target.

Table A-7 Required number of replications for each facility for α = 0.005 and γ=0.005

	Villach	Regensburg	Kulim
Existing model	5	9	7
Updated model	8	9	9

As we already have data of 10 replications, we decide to use this number of replications. We can therefore conclude, that our results lie in a confidence interval >99.5% with a relative error <0.5%.

Appendix L: Accuracy

Figure A-7 illustrates where we made changes to the existing simulation model based on the relationship chart we introduced in Section 2.2.5 (Figure 2-8).

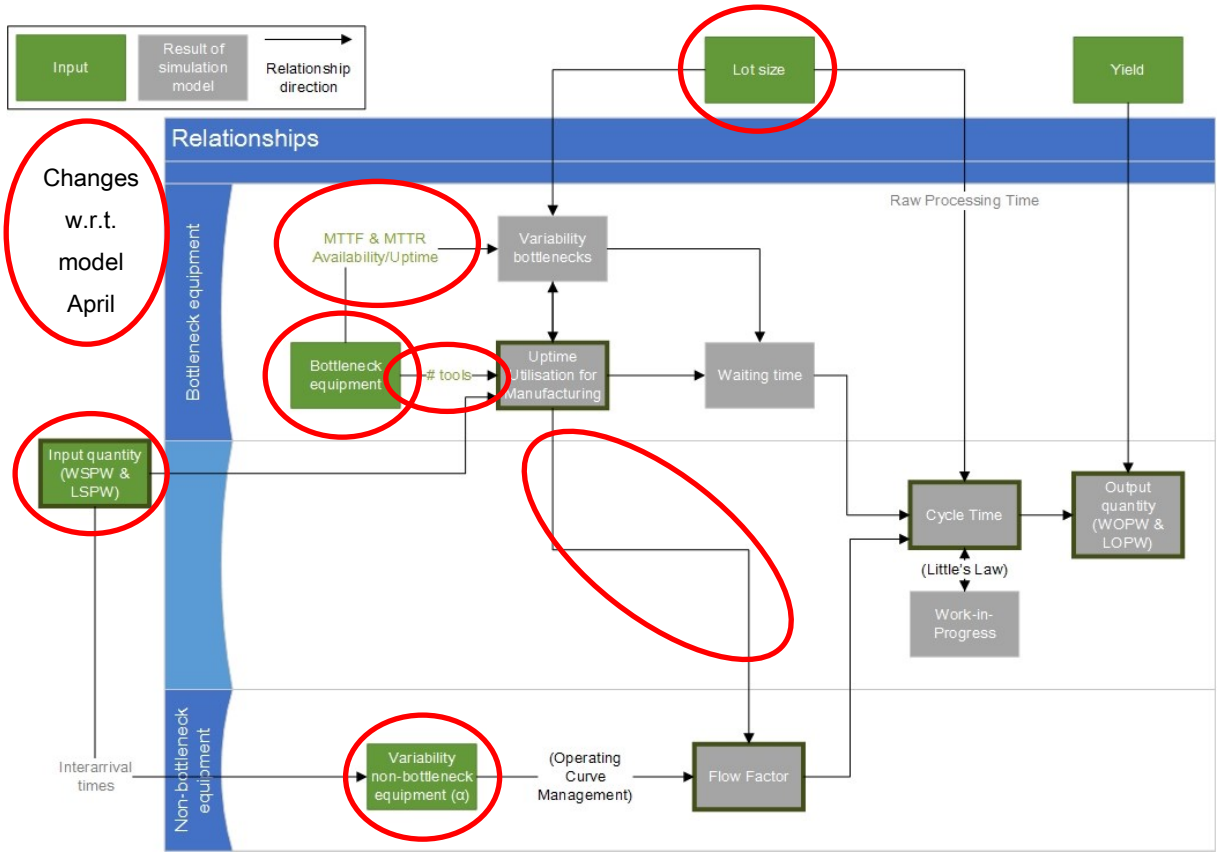


Figure A-0-7 Relationship chart with the improved areas marked

These changes explain the improved overall accuracy measure Δ (Table 4-2 (Section 4.2.3)) and the improved accuracy measure Δ for each facility (Table A-8) with respect to the existing simulation model. We explain this for each KPI separately:

WSPW

The overall Δ and the Δ per facility for the WSPW of the updated simulation model improved in comparison with the accuracy of the existing simulation model (Table 4-2 & Table A-8). This is due to the fact that most of the missing data for the WSPW, which were left out in the existing model, have been recovered. If we look at the relationship chart (Figure A-7), the positive change of the input data has an impact on the rest of the simulation model.

WOPW

Even though the overall Δ and the Δ per facility for the WOPW shown by the tables (Table 4-2 & Table A-8) did not improve in comparison with the existing simulation model, we cannot say if the simulation model performs worse in this case. As wafer losses, which are not included in the yield, are not taken into account in the simulation model, we can explain the deviation from the historical value. However, we do not know the real impact. Furthermore, the increased (positive) value can be explained by the fact that we completed the input data.

LSPW

We can explain the difference between the overall Δ and the Δ per facility for the LSPW and the Δs for the WSPW (Table 4-2 & Table A-8), by the fact that the routes of the lots (so also the lithography steps) are aggregated on process line level for the simulation model. However, the historical LSPW is determined per product and its individual route.

LOPW

The same reasoning can be used for the LOPW as for the WOPW. Moreover, in the completed input data the routes of the lots (so also the lithography steps) are aggregated on process line level. However, the historical LSPW is determined per product and its individual route.

CT

The accuracy of the CT increased greatly in the updated simulation model in comparison with the existing simulation model (Table 4-2 & Table A-8). This is due to the many changes made (Figure 30) that influence the CT, such as the facility-specific lot sizes, the facility-specific and less aggregated bottlenecks, and the variability (α).

Utilisation

The accuracy of the utilisation improved in the updated simulation model existing simulation model in comparison with the accuracy of in the existing simulation model, both overall and per facility (Table 4-2 & Table A-8). The utilisation is influenced by the bottlenecks and their parameterisation (Figure A-7). This is where we made changes and it had a positive impact.

FF

The accuracy of the FF improved as well, caused by a lot of refinements we made to the existing simulation model (Table 4-2 & Table A-8), such as the process group specific

variability (α), the facility-specific utilisation used for the calculation, the less aggregated bottlenecks and their parameterisation, and the improved lot sizes (Figure A-7).

Table A-8 Accuracy per KPI per facility per KPI for the existing model and the updated model

Simulation model		WSPW	WOPW	LSPW	LOPW	CT	Utilisation	FF
Villach	Existing	-2,3%	1,1%	-0,6%	2,8%	-11,8%	-2,4%	-4,1%
	Updated	-0,4%	2,9%	2,3%	5,8%	6,6%	0,6%	6,2%
Regensburg	Existing	-3,6%	1,9%	-1,2%	5,1%	-17,5%	-1,5%	-2,1%
	Updated	-0,7%	2,0%	1,4%	5,2%	-0,3%	-0,5%	-1,8%
Kulim	Existing	-0,3%	4,4%	-0,6%	5,3%	-36,4%	-5,0%	-8,3%
	Updated	0,0%	6,1%	-0,4%	7,1%	-3,0%	-3,0%	2,8%

Appendix M: Results contingency experiments

Table A-9 lists the results averaged over the 5 replications for each of the 180 factor level combinations. The base case is made bold.

Table A-9 Experiment results

Experiment	Average Die Bank reach (days)	Average DC reach (days)	Theoretical utilisation (%)	Costs (%) vs. base case	SLα (%)	SLβ (%)	SLγ (%)
1	0	0	75%	99,44%	53,4%	86,6%	21,7%
2	7	0	75%	99,46%	61,5%	87,2%	45,7%
3	14	0	75%	99,48%	73,5%	91,1%	62,7%
4	21	0	75%	99,50%	79,1%	93,2%	69,3%
5	28	0	75%	99,57%	82,5%	94,3%	70,9%
6	35	0	75%	99,60%	83,3%	94,4%	71,0%
7	0	7	75%	99,47%	64,9%	87,1%	56,2%
8	7	7	75%	99,51%	77,4%	90,5%	69,5%
9	14	7	75%	99,57%	85,2%	93,5%	78,5%
10	21	7	75%	99,61%	89,1%	94,4%	80,3%
11	28	7	75%	99,64%	89,5%	94,6%	79,9%
12	35	7	75%	99,67%	89,3%	94,5%	80,1%
13	0	14	75%	99,53%	79,9%	91,6%	77,0%
14	7	14	75%	99,60%	88,1%	94,7%	86,5%
15	14	14	75%	99,63%	91,2%	95,7%	87,9%
16	21	14	75%	99,69%	91,6%	95,7%	88,0%
17	28	14	75%	99,71%	91,4%	95,6%	87,7%
18	35	14	75%	99,74%	91,7%	95,8%	88,0%
19	0	21	75%	99,62%	92,4%	97,2%	94,4%
20	7	21	75%	99,67%	94,7%	97,7%	95,1%
21	14	21	75%	99,72%	95,1%	97,8%	95,3%
22	21	21	75%	99,77%	95,0%	97,8%	95,1%
23	28	21	75%	99,79%	95,0%	97,8%	95,2%
24	35	21	75%	99,83%	95,0%	97,8%	95,2%
25	0	28	75%	99,69%	98,5%	99,8%	99,8%
26	7	28	75%	99,77%	98,7%	99,8%	99,8%
27	14	28	75%	99,82%	98,8%	99,8%	99,8%
28	21	28	75%	99,83%	98,8%	99,8%	99,7%
29	28	28	75%	99,87%	98,8%	99,8%	99,8%

Experiment	Average Die Bank reach (days)	Average DC reach (days)	Theoretical utilisation (%)	Costs (%) vs. base case	SLα (%)	SLβ (%)	SLγ (%)
30	35	28	75%	99,91%	99,0%	99,8%	99,8%
31	0	35	75%	99,78%	100,0%	100,0%	100,0%
32	7	35	75%	99,84%	100,0%	100,0%	100,0%
33	14	35	75%	99,91%	100,0%	100,0%	100,0%
34	21	35	75%	99,95%	100,0%	100,0%	100,0%
35	28	35	75%	99,99%	100,0%	100,0%	100,0%
36	**35**	**35**	**75%**	**100,00%**	**100,0%**	**100,0%**	**100,0%**
37	0	0	80%	93,42%	57,8%	85,5%	30,1%
38	7	0	80%	93,37%	61,3%	86,5%	36,6%
39	14	0	80%	93,38%	62,8%	87,4%	45,4%
40	21	0	80%	93,48%	70,7%	89,4%	56,4%
41	28	0	80%	93,45%	74,8%	91,6%	64,2%
42	35	0	80%	93,58%	79,1%	92,4%	68,3%
43	0	7	80%	93,30%	63,8%	86,9%	37,6%
44	7	7	80%	93,46%	66,8%	86,6%	47,4%
45	14	7	80%	93,51%	74,4%	88,9%	62,5%
46	21	7	80%	93,61%	81,3%	90,8%	70,5%
47	28	7	80%	93,62%	85,2%	92,6%	75,3%
48	35	7	80%	93,66%	87,6%	93,3%	77,4%
49	0	14	80%	93,54%	66,3%	86,3%	49,4%
50	7	14	80%	93,60%	75,5%	88,7%	63,7%
51	14	14	80%	93,62%	81,2%	90,7%	72,2%
52	21	14	80%	93,64%	85,0%	92,4%	76,9%
53	28	14	80%	93,58%	88,0%	94,5%	78,5%
54	35	14	80%	93,64%	88,5%	94,5%	78,7%
55	0	21	80%	93,55%	75,7%	90,1%	69,4%
56	7	21	80%	93,57%	83,3%	92,5%	79,2%
57	14	21	80%	93,58%	86,9%	94,3%	83,3%
58	21	21	80%	93,66%	89,3%	94,7%	84,5%
59	28	21	80%	93,70%	89,4%	94,8%	84,0%
60	35	21	80%	93,79%	89,4%	94,5%	84,7%
61	0	28	80%	93,58%	87,3%	94,6%	85,5%
62	7	28	80%	93,69%	90,9%	95,4%	89,5%
63	14	28	80%	93,71%	92,4%	96,0%	91,0%

Experiment	Average Die Bank reach (days)	Average DC reach (days)	Theoretical utilisation (%)	Costs (%) vs. base case	SLα (%)	SLβ (%)	SLγ (%)
64	21	28	80%	93,75%	93,0%	96,3%	91,0%
65	28	28	80%	93,87%	93,0%	95,7%	90,2%
66	35	28	80%	93,82%	93,0%	96,5%	90,8%
67	0	35	80%	93,71%	93,4%	96,9%	93,6%
68	7	35	80%	93,78%	95,8%	98,0%	95,7%
69	14	35	80%	93,78%	96,2%	98,3%	96,6%
70	21	35	80%	93,83%	96,2%	98,3%	96,8%
71	28	35	80%	93,88%	96,2%	98,3%	96,7%
72	35	35	80%	93,94%	95,9%	98,0%	96,1%
73	0	0	85%	87,93%	48,1%	84,5%	26,1%
74	7	0	85%	87,95%	57,0%	85,2%	33,6%
75	14	0	85%	88,01%	57,7%	85,8%	35,6%
76	21	0	85%	88,00%	59,6%	86,0%	39,6%
77	28	0	85%	88,05%	63,5%	87,8%	47,3%
78	35	0	85%	88,06%	68,4%	89,6%	55,2%
79	0	7	85%	87,95%	59,7%	85,4%	33,6%
80	7	7	85%	88,01%	60,5%	85,6%	36,2%
81	14	7	85%	88,06%	63,1%	86,3%	41,9%
82	21	7	85%	88,10%	67,3%	88,0%	50,7%
83	28	7	85%	88,08%	72,2%	89,3%	59,0%
84	35	7	85%	88,14%	77,8%	91,1%	66,4%
85	0	14	85%	88,06%	60,9%	85,8%	36,0%
86	7	14	85%	88,08%	62,1%	86,2%	41,2%
87	14	14	85%	88,09%	67,4%	87,8%	51,7%
88	21	14	85%	88,10%	73,2%	89,7%	60,7%
89	28	14	85%	88,11%	77,9%	91,2%	67,4%
90	35	14	85%	88,21%	81,7%	92,5%	72,1%
91	0	21	85%	88,09%	62,7%	86,3%	42,0%
92	7	21	85%	88,10%	68,3%	88,0%	52,3%
93	14	21	85%	88,17%	73,5%	89,9%	60,5%
94	21	21	85%	88,20%	78,4%	91,2%	67,5%
95	28	21	85%	88,24%	81,9%	92,6%	73,1%
96	35	21	85%	88,28%	85,0%	93,7%	76,2%
97	0	28	85%	88,12%	68,0%	88,1%	53,4%

Experiment	Average Die Bank reach (days)	Average DC reach (days)	Theoretical utilisation (%)	Costs (%) vs. base case	SLα (%)	SLβ (%)	SLγ (%)
98	7	28	85%	88,16%	72,7%	89,6%	62,1%
99	14	28	85%	88,22%	78,6%	91,5%	69,3%
100	21	28	85%	88,24%	82,3%	92,6%	74,7%
101	28	28	85%	88,31%	85,0%	93,8%	78,1%
102	35	28	85%	88,36%	87,0%	94,4%	79,4%
103	0	35	85%	88,18%	75,3%	90,7%	66,8%
104	7	35	85%	88,22%	80,0%	92,2%	73,8%
105	14	35	85%	88,36%	84,7%	93,7%	78,8%
106	21	35	85%	88,34%	86,4%	94,6%	81,8%
107	28	35	85%	88,39%	88,5%	95,0%	83,4%
108	35	35	85%	88,46%	88,6%	95,6%	83,6%
109	0	0	90%	83,27%	53,5%	83,6%	24,7%
110	7	0	90%	83,26%	55,5%	83,5%	33,6%
111	14	0	90%	83,38%	57,5%	84,2%	34,9%
112	21	0	90%	83,36%	58,8%	84,7%	38,1%
113	28	0	90%	83,39%	62,4%	86,2%	42,1%
114	35	0	90%	83,38%	63,7%	86,9%	47,0%
115	0	7	90%	83,29%	59,2%	83,8%	33,7%
116	7	7	90%	83,33%	59,8%	83,9%	35,3%
117	14	7	90%	83,34%	61,0%	84,5%	38,9%
118	21	7	90%	83,42%	65,2%	85,9%	43,6%
119	28	7	90%	83,44%	68,4%	87,2%	49,7%
120	35	7	90%	83,45%	72,8%	88,6%	56,3%
121	0	14	90%	83,35%	59,7%	83,9%	35,4%
122	7	14	90%	83,39%	62,5%	84,9%	39,2%
123	14	14	90%	83,44%	64,9%	85,9%	43,9%
124	21	14	90%	83,46%	68,6%	87,5%	50,7%
125	28	14	90%	83,51%	72,4%	88,5%	56,2%
126	35	14	90%	83,50%	76,3%	89,9%	63,1%
127	0	21	90%	83,39%	62,4%	84,9%	39,6%
128	7	21	90%	83,46%	65,0%	86,0%	44,3%
129	14	21	90%	83,51%	68,7%	87,2%	50,3%
130	21	21	90%	83,56%	73,1%	88,7%	56,6%
131	28	21	90%	83,49%	75,5%	89,7%	63,3%

Experiment	Average Die Bank reach (days)	Average DC reach (days)	Theoretical utilisation (%)	Costs (%) vs. base case	SLα (%)	SLβ (%)	SLγ (%)
132	35	21	90%	83,56%	79,7%	91,3%	68,7%
133	0	28	90%	83,48%	65,0%	86,1%	44,5%
134	7	28	90%	83,54%	69,4%	87,6%	51,0%
135	14	28	90%	83,55%	72,9%	88,7%	57,1%
136	21	28	90%	83,56%	75,9%	89,8%	63,2%
137	28	28	90%	83,63%	79,8%	91,6%	68,8%
138	35	28	90%	83,64%	83,4%	92,6%	73,0%
139	0	35	90%	83,55%	69,2%	87,5%	51,6%
140	7	35	90%	83,52%	72,0%	88,4%	58,7%
141	14	35	90%	83,64%	76,9%	90,0%	64,6%
142	21	35	90%	83,64%	80,3%	91,5%	70,5%
143	28	35	90%	83,73%	84,2%	93,0%	74,1%
144	35	35	90%	83,74%	85,4%	93,5%	76,8%
145	0	0	95%	79,19%	49,7%	81,6%	18,5%
146	7	0	95%	79,20%	56,2%	82,4%	35,5%
147	14	0	95%	79,22%	60,3%	83,3%	41,2%
148	21	0	95%	79,26%	62,3%	84,3%	44,7%
149	28	0	95%	79,23%	64,8%	85,9%	48,1%
150	35	0	95%	79,31%	67,1%	86,9%	51,6%
151	0	7	95%	79,21%	55,5%	82,4%	37,7%
152	7	7	95%	79,24%	63,5%	83,5%	42,8%
153	14	7	95%	79,29%	64,8%	84,3%	46,5%
154	21	7	95%	79,30%	69,1%	85,7%	51,0%
155	28	7	95%	79,33%	71,8%	87,0%	55,5%
156	35	7	95%	79,35%	75,0%	88,5%	61,0%
157	0	14	95%	79,27%	64,2%	83,4%	42,8%
158	7	14	95%	79,30%	66,1%	84,3%	47,0%
159	14	14	95%	79,32%	69,5%	85,6%	51,0%
160	21	14	95%	79,32%	71,6%	86,9%	56,1%
161	28	14	95%	79,38%	74,9%	88,5%	61,8%
162	35	14	95%	79,40%	79,2%	90,3%	66,6%
163	0	21	95%	79,31%	66,5%	84,7%	47,4%
164	7	21	95%	79,35%	69,2%	85,7%	51,3%
165	14	21	95%	79,41%	72,9%	86,9%	56,1%

Experiment	Average Die Bank reach (days)	Average DC reach (days)	Theoretical utilisation (%)	Costs (%) vs. base case	SLα (%)	SLβ (%)	SLγ (%)
166	21	21	95%	79,44%	75,3%	88,6%	61,8%
167	28	21	95%	79,47%	79,4%	90,2%	66,6%
168	35	21	95%	79,49%	81,3%	91,0%	69,9%
169	0	28	95%	79,36%	69,0%	85,7%	51,7%
170	7	28	95%	79,43%	72,0%	87,2%	56,3%
171	14	28	95%	79,45%	75,9%	88,8%	62,4%
172	21	28	95%	79,48%	78,2%	90,1%	66,6%
173	28	28	95%	79,51%	81,2%	91,1%	70,1%
174	35	28	95%	79,55%	82,9%	92,0%	72,1%
175	0	35	95%	79,43%	72,1%	87,3%	56,9%
176	7	35	95%	79,46%	75,6%	88,9%	62,5%
177	14	35	95%	79,56%	79,9%	90,3%	67,2%
178	21	35	95%	79,54%	81,6%	91,5%	69,9%
179	28	35	95%	79,57%	83,5%	92,0%	72,5%
180	35	35	95%	79,63%	85,3%	92,6%	74,4%

Figures A-8, A-9, and A-10 show how the different service level types relate to each factor level combination. All the curves follow the same pattern. However, the distribution in the graph of the SL^{γ} occupies a larger range than the other service level types. The graph visualising the distribution for the SL^{β} occupies the smallest range.

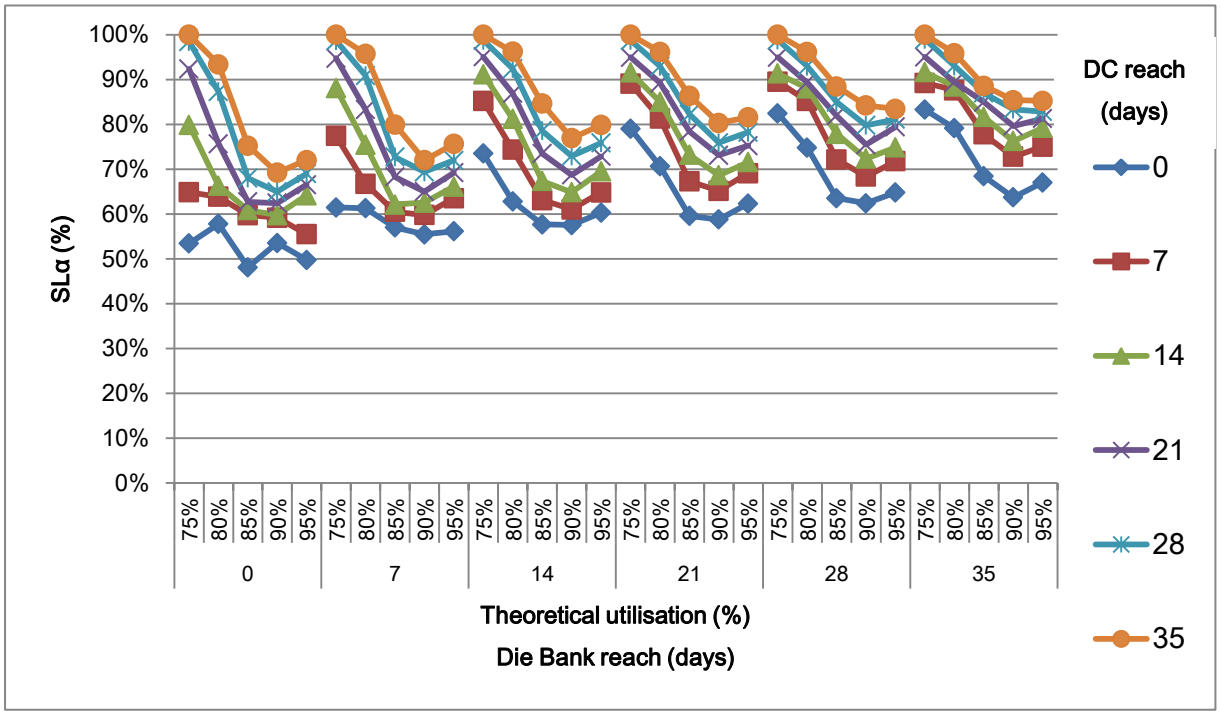


Figure A-8 Relationship between SL^{α} and the factor level combinations

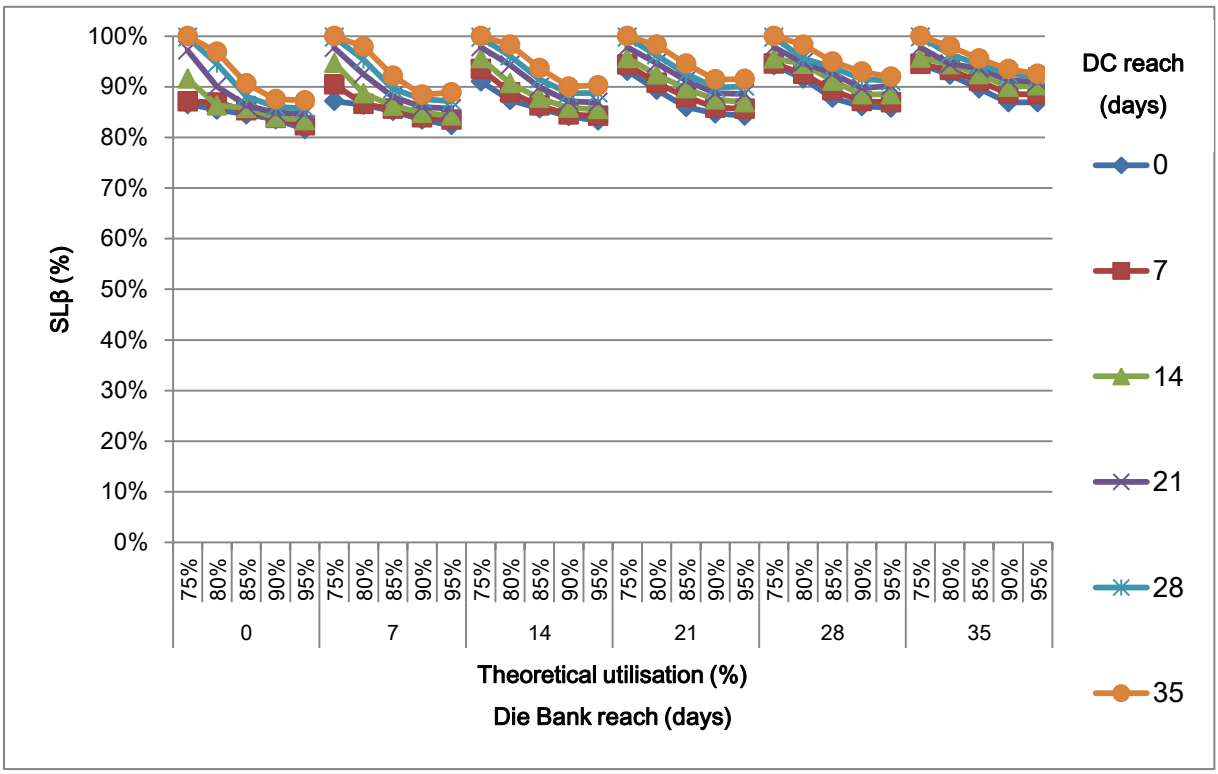


Figure A-9 Relationship between SL^{β} and the factor level combinations

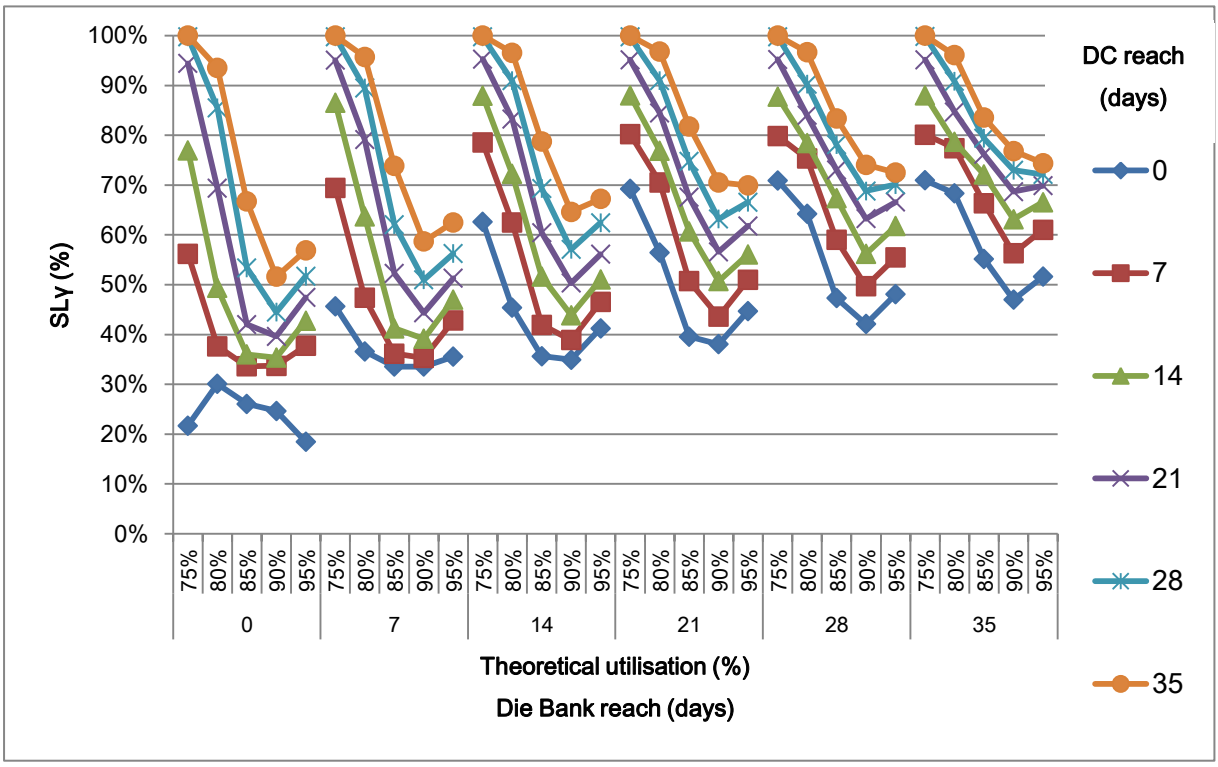


Figure A-10 Relationship between SLʸ and the factor level combinations

Figure A-11 confirms our expected cause by an example with a die bank reach and DC reach of 21 days. The figure shows that the backorder level (including backorders of previous periods) which is used for the SLʸ calculation is much lower and more constant for 95% utilisation than for 90% utilisation at which the backorder level keeps increasing.

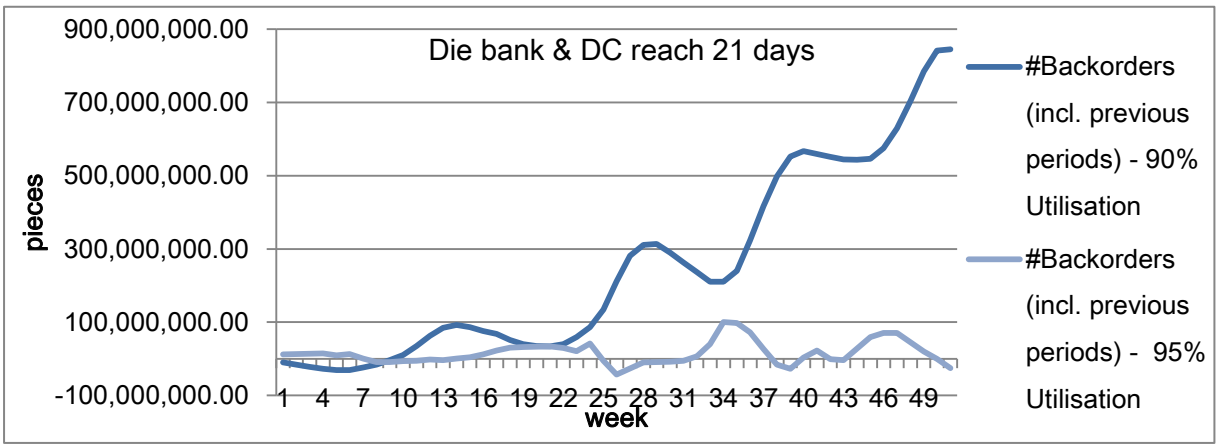


Figure A-11 Backorder level (incl. backorders of previous weeks) for 90% and 95% utilisation

Figure A-12 shows the die bank and DC stock levels for 90% and 95% utilisation (at a die bank and DC reach of 21 days). The die bank stock level for 95% utilisation is lower than the die bank stock level for 90% utilisation. However, for the DC stock level it is the other way round.

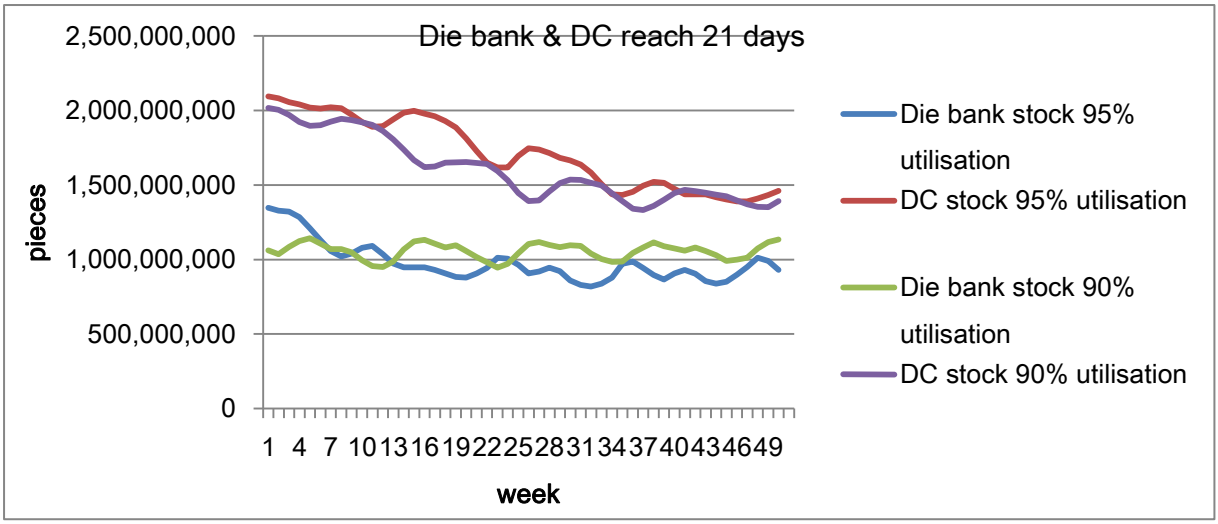


Figure A-12 Stock levels at 90% and 95% utilisation