Universität der Bundeswehr München

Online Simulation in Semiconductor Manufacturing

Daniel Noack

Vollständiger Abdruck der von der Fakultät für Informatik der Universität der Bundeswehr München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation.

Gutachter:
1. Prof. Dr. rer. nat. Oliver Rose
2. Priv.-Doz. Dr.-Ing. Gerald Weigert (TU Dresden)

Die Dissertation wurde am 14.05.2012 bei der Universität der Bundeswehr München eingereicht und durch die Fakultät für Informatik am 06.09.2012 angenommen. Die mündliche Prüfung fand am 25.09.2012 statt.

Abstract

In semiconductor manufacturing discrete event simulation systems are quite established to support multiple planning decisions. During the recent years, the productivity is increasing by using simulation methods. The motivation for this thesis is to use online simulation not only for planning decisions, but also for a wide range of operational decisions. Therefore an integrated online simulation system for short term forecasting has been developed. The production environment is a mature high mix logic wafer fab. It has been selected because of its vast potential for performance improvement. In this thesis several aspects of online simulation will be addressed:

The first aspect is the **implementation** of an online simulation system in semiconductor manufacturing. The general problem is to achieve a high speed, a high level of detail, and a high forecast accuracy. To resolve these problems, an online simulation system has been created. The simulation model has a high level of detail. It is created automatically from underling fab data. To create such a simulation model from fab data, additional problems related to the underlying data arise. The major parts are the data access, the data integration, and the data quality. These problems have been solved by using an integrated data model with several data extraction, data transformation, and data cleaning steps.

The second aspect is related to the **accuracy** of online simulation. The overall problem is to increase the forecast horizon, increase the level of detail of the forecast and reduce the forecast error. To provide useful forecast results, the simulation model contains a high level of modeling details and a proper initialization. The influences on the forecast quality will be analyzed. The results show that the simulation forecast accuracy achieves good quality to predict future fab performance.

The last aspect is to find ways to use simulation forecast results to improve the fab performance. Numerous **applications** have been identified. For each application a description is available. It contains the requirements of such a forecast, the decision variables, and background information. An application example shows, where a performance problem exists and how online simulation is able to resolve it.

To further enhance the **real time** capability of online simulation, a major part is to investigate new ways to connect the simulation model with the wafer fab. For fab driven simulation, the simulation model and the real wafer fab run concurrently. The wafer fab provides several events to update the simulation during runtime. So the model is always synchronized with the real fab. It becomes possible to start a simulation run in real time. There is no further delay for data extraction, data transformation and model creation. A prototype for a single work center has been implemented to show the feasibility.

Keywords
Online Simulation, Semiconductor Manufacturing, Forecasting, Operational Decision Support

Acknowledgements

First I would like to thank Prof. Oliver Rose to give me the opportunity to do my research in my desired field of interest. He has provided excellent research conditions and a high degree of freedom during my research. Without his support, this PhD thesis would not have been possible.

A big thank you goes to Peter Lendermann and Gan Boon Ping. They gave me the opportunity to participate in an industry related online simulation project. They have organized a lot to let this project happen. They invested a lot of time and money to push online simulation forward. They also provided me with crucial help to combine academia and industry in my PhD.

Many thanks go to Wolfgang Scholl, for his guidance in a real industrial environment. He provided a lot of background information in underground and industrial environments. The differences between industry and academia have become obvious. So many thanks for the excursus into practice.

I also would like to thank my parents. During my studies, they were there to help me focus on those things which are important. They gave me a different perspective, apart the details and difficulties of this work. They provided stability and safety during that time.

1 Introduction

The introduction chapter provides an overview of the production environment. It presents the motivation and challenges of this thesis. A clear picture of the four main objectives is given.

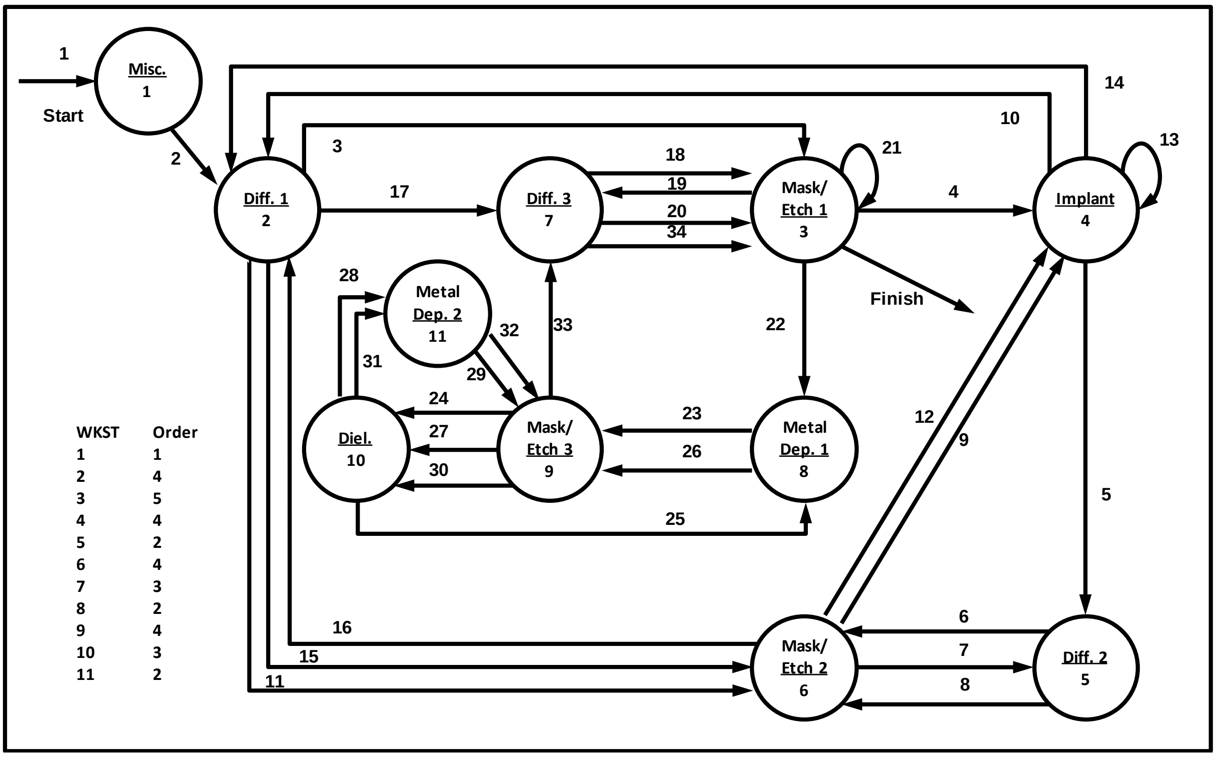
1.1 Production Environment

The production environment for this thesis is semiconductor manufacturing. The online simulation system has been implemented in a large high mix wafer fab for logic devices.

Semiconductor manufacturing is a very complex manufacturing environment (Atherton and Atherton 1995). Modern manufacturing systems are designed for mass production. It is hardly achievable to organize the manufacturing process as a flow shop for a front end wafer fab out of multiple reasons:

- High semiconductor equipment cost
- High product mix
- Reentrant process flows
- Unpredictable equipment downs

Unfortunately the job shop principle which is used, contradicts the principle of mass production. So, the fab layout is not organized according to the lot flow. Instead, typically resources with similar processes are located together as a work center. The lots, which contain several wafers, travel from one work center to the next, according to their route (Figure 1).



WKST	Order
1	1
2	4
3	5
4	4
5	2
6	4
7	3
8	2
9	4
10	3
11	2

Figure 1: Fab graph (Atherton and Atherton 1995)

Beside the basic characteristics additional problems increase uncertainty and interfere with the continuous process flow. Following reasons exist for high uncertainty:

- Rework
- Sampling
- Hold
- Variable process times
- Lot priorities
- Time constraints
- Setup
- Mix of single wafer and batching tools
- Dedicated equipment

It is obvious that many disturbances but also the nature of the fab contribute to such a high complexity (Hopp and Spearman 2000, Robinson 1998). Complexity is an even bigger problem because of the large size of the fab. Several hundred pieces of equipment, different process flows, and thousands of lots interact in the wafer fab.

Another fact is that the costs for semiconductor fabs are huge. According to McGregor (2007) the cost for a semiconductor fab is estimated at around 5 billion US Dollar. Because of high investment cost and a vast performance improvement potential, several methods have been used to increase productivity.

1.2 Motivation

The major motivation aspect is the performance improvement potential in semiconductor manufacturing. In a complex and dynamic manufacturing environment it is nearly impossible to make optimal operational decisions without in-depth knowledge of a wafer fab behavior. Multiple effects of a dynamic manufacturing environment cannot be addressed exactly with static data analysis methods and personnel experiences. If disturbances are not well managed, the fab operation will become very unpredictable. Unpredictability typically results in a decrease in fab productivity and high production cost. It affects the company competitiveness on the global market place. By using simulation for performance prediction, multiple operational decisions can be improved considering future fab conditions. The decision process will be rather proactive instead of reactive. Using simulation for manufacturing operations in a front end semiconductor fab is highly effective for this complex environment.

The next motivation aspect is that simulation based methods are already well known, especially for planning purposes. Simulation is highly valuable due to the possibility to rely on a model for system analysis instead of a real system. It becomes highly useful due to the capability of what-if scenarios and optimization. In the context of semiconductor manufacturing, simulation has been used for planning decisions for many years. The motivation is to make use of discrete event simulation not only for planning but also for a wide range of operational decisions. Discrete event simulation results need to be available directly on the factory floor, to improve factory performance further.

Another source for the motivation is that online simulation is a big challenge because the overall problem is difficult to solve. The general problem is to provide simulation based results for operational decision support in real time. This “real time” capability is a central problem in semiconductor manufacturing because the problem size is immense and of very complex nature. This complexity and large problem size are a direct contradiction to the real time requirement.

The motivation for this thesis is a combination of all three elements.

- The advantages of discrete event simulation method
- The performance improvement potentials in semiconductor manufacturing
- The challenge to implement a simulation based system with real time capability

1.3 Challenges

Creating an online simulation system for operational management in a front end semiconductor manufacturing facility is a grand challenge. According to Crosbie (2010) and his definition of a grand challenge, it means that short term simulation…

- …must be hard to solve
- …must not be known to be unsolvable
- …must have economical /social impact

J. Fowler pointed out that short term simulation has much potential for economic impact (Fujimoto et al. 2002). He describes that "electronics industry recently surpassed the automotive industry to become the largest basic industry in the world after agriculture". Increasing productivity improvements come from "wafer size changes", "devices shrinks", "yield improvements", and "factory and equipment efficiency improvements". Fab and equipment efficiency improvements offer the highest potential for future gains. Furthermore he stated that "operational modeling and simulation offer a way to determine areas that will lead to significant enterprise, factory and equipment efficiency improvements."

Fowler et al. (2002) pointed out that short term simulation is a challenge. It is necessary to provide "real-time simulation-based problem solving capability". It is required to do "what-if analysis at any time". Also the "Factory status constantly changes" which requires that a "persistent model constantly updated from manufacturing system". Increasing execution speed is required because "the time needed to collect and synthesize the required information/data and the time required to do the experimentation are simply too long."

H. Szczerbicka, also pointed out that "On-line simulation is a new technology for on-line planning and controlling of systems, which needs further research" (Fujimoto et al. 2002). Therefore there are several aspects in the range of interest. Research examples are automated model validation, fast experiment generation, performance prediction, online analysis for control policies, and speed up of execution time.

1.4 Objectives

The overall objective of this thesis is to increase the performance of the manufacturing system by using online simulation. For online simulation multiple aspects are in the range of interest for the research area. For this thesis the following key objectives have been selected:

- Describe the implementation methodology of online simulation
- Achieve high accuracy
- Achieve real time capability
- Identify applications for online simulation forecast results

The online simulation system is implemented in a real wafer fab. The implementation sections provide real world examples for obstacles, but also for solution approaches in online simulation. In this context the word implementation is used as a synonym for the whole developments context, including the concept, design and individual results.

The high accuracy is important because the results on operational level affect only a small part of the fab, for example one work center. A simulation model needs to reflect this work center with sufficient accuracy to provide useful forecast results.

The real time capability is necessary to provide results, before they are obsolete. For online simulation the timing requirements are very high, depending on the particular operation. The high computation time for data transformation, model initiation, and simulation is always an issue.

The identification of applications is another important aspect. It demonstrates the capability to use the forecast results. A simulation forecast without an integrated process to use the results for performance improvements does not have much value.

1.4.1 Implementation

The first objective is to show the methodology to develop an online simulation system for short term forecasting in a complex industrial environment. It is divided into the simulation modeling part and the data modeling part. The objective of the simulation modeling part is to show how to provide forecast results very fast, with a high level of detail, and high forecast accuracy. The second part is referring to the underlying **data** of online simulation. Most important problems are related to data integration, data quality, and the speed to provide all data in time. The objective is to list the data problems and to find intelligent solutions to extract data, to transform data, and to load data. It is necessary to meet the data requirements for online simulation. The feasibility and accuracy of online simulation highly depends on the data input. The following questions will be addressed:

- How is it possible to develop and implement online simulation for a full fab model?
- What is a good way to handle data integration?

1.4.2 Accuracy

The major objective is to reach a high modeling accuracy. The modeling accuracy depends on the level of detail, the simulation forecast horizon, and the deviation from reality (Figure 2). The higher the level of detail, the higher is the deviation from reality. Also the higher the forecast time horizon, the higher is the uncertainty and deviation from reality. The forecast quality needs to be high to address a wide range of operational problems. Therefore the objective is to maximize the forecast level of detail, maximize the time horizon, and reduce the forecasting error. One objective is to reflect the forecast quality in a single overview chart, like seen in see Figure 2.

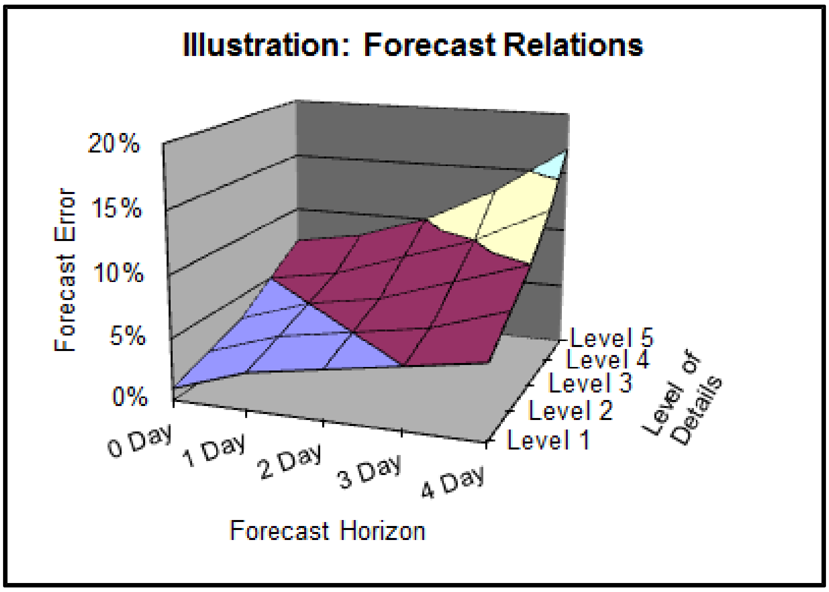


Figure 2: Illustration of model accuracy relations

To increase the modeling accuracy it is necessary to do further analysis. One objective is to analyze the influence of modeling features on the forecast accuracy. A key problem is how to deal with uncertainty. Uncertainties like unscheduled tool downs, sampling, and variable process times are hard to handle. But uncertainty is critical to the accuracy of the simulation model.

Another objective is to reduce the warm-up period and increase the model accuracy at simulation start. It is necessary to figure out what are the reasons for the deviation of reality and simulation during the short time period right after simulation start. Only when these warm-up effects are known, it is possible to reduce them efficiently. One objective is to find out ways to enhance the quality of the model initialization and to reduce the warm-up effects. Regarding the modeling accuracy following questions will be addressed:

- What is the achieved accuracy of the online simulation model?
- What are the reasons for the model deviation at start-up (warm-up period)?
- How is it possible to increase initialization quality to reduce the deviation at warm-up period?
- How much does a detailed the model initialization affect the accuracy of the forecast?
- What is the influence of stochastic effects? Is a deterministic approach feasible?

1.4.3 Real Time Capability

To further enhance the real time capability of online simulation one objective is to provide the proof of concept of the fab driven simulation approach. It is necessary to investigate new ways to couple the wafer fab with the simulation model. The objective is to implement a prototype, to demonstrate that a small model is synchronized with fab data. The model runs concurrently with the wafer fab and it is possible to start the simulation run immediately. Following questions will be addressed:

- What is the limitation regarding the real time capability of the current implementation?
- How is it possible to overcome such a limitation?

1.4.4 Applications

First it is necessary to define what an application for online simulation is. The term application is used to capture how exactly the forecast results are used to improve the fab performance. It contains one or multiple decision variables, combined with a decision process. The value of such decision variables can be changed. Changes of decision variables affect the factory performance. For every application a decision process exists. A person or a mechanism is available, which change the value of the decision variable. For every decision, different input data is required. For each application several constraints also exist, for example time constraints between the decision itself and the time when it takes effect in the fab.

For online simulation, the objective is to identify and describe numerous operational applications. The general question is how simulation forecast results will be used, to improve the performance of the wafer fab. It is necessary, to identify the decision variable that can be changed. The objective is to list many applications, provide detailed description, and reveal useful background information. It is necessary to describe the required application input information and the application constraints, to derive the requirements of an online simulation forecast. The objective is to figure out, if an application is compatible with the simulation forecasting approach. So every operational planning problem has its specific requirements for the accuracy, the forecast horizon, and the level of detail. The objective is to figure out if the forecast quality is valid to install such an application in the daily decision process. The main questions are:

- Which online simulation applications exist to increase fab performance?
- Does simulation forecasting satisfy the requirements of such applications?

1.5 Structure of this Thesis

Chapter 2 contains a short description of the semiconductor manufacturing environment. An explanation is available why online simulation is such a big challenge. It contains the description of problems which need to be addressed to realize online simulation.

The current state of the art of online simulation is available in chapter 3. It also presents an overview of related work areas. It starts from a general perspective up to the detailed level. Topics are operational decision support in semiconductor manufacturing, modeling, optimization and forecasting. This chapter provides the reasons why the simulation method is used to solve operational problems. On a detailed level, literature sources are available for data modeling, simulation modeling accuracy, and applications to use forecast results.

Chapter 4 and 5 contain the system descriptions for online simulation. It consists of the data modeling part and the simulation modeling part. For the data model, the requirements, the concept and the implementation and results are described. The focus of the data modeling part is the data integration. The simulation modeling part consists of requirements, the conceptual model and the executable model part. The focus is to describe how to model single elements of the wafer fab, to mimic the behavior of the whole wafer fab.

The achievements, regarding the modeling accuracy, are presented in chapter 6. It contains a critical evaluation of the simulation results and the effect of simulation elements on forecast accuracy. A major part is the analysis of simulation accuracy at simulation start.

Chapter 7 presents the underlying applications of online simulation. Those applications improve the fab performance by using forecast results. This list describes those applications in detail. One part is a scenario which shows how exactly a simulation forecast result is able to improve the fab performance.

Chapter 8 contains the fab driven simulation part. It shows how it is possible to continuously update a small simulation model while the simulation is running. This part includes the requirements, the concept, implementation and results. The focus is on the proof of concept to show that fab driven simulation is working in a real fab environment.

The conclusions in chapter 9 highlight the achievements of this thesis and the contribution to science. These achievements are the implementation of online simulation in an industrial environment, the analysis of simulation accuracy, the applications to increase the fab performance, and the fab driven simulation.

2 Problem Description

Within the pervious chapter, the objectives of online simulation have become clear. In order to reach these objectives, multiple problems exist which need to be solved. This chapter describes these problems in detail. It contains the problem description of the four major parts of this thesis:

- Implementation problems for the data model and the simulation model
- Problems related to the simulation modeling accuracy
- Problem to achieve real time capability
- Problems to identify applications for online simulation

2.1 Implementation of Online Simulation

A major problem is the development of an online simulation system for short term forecasting in a real manufacturing environment. This is a complex task. Many elements exist which become part of the simulation model and part of the whole software solution. The complexity is very high, because many elements influence each other. There are also many disturbances in the wafer fab which have an effect on the forecast quality. The combination of requirements is hard to solve. It is necessary to provide high forecast accuracy, almost in real time, dealing with a large variety of problems, and the high complexity. The creation of online simulation comes with several problems. All elements are part of the software engineering process:

- Analysis of requirements
- Concept
- Implementation
- Validation

2.1.1 Requirements

The first task is to define the requirements of online simulation. These requirements define the targets of the concept and implementation part. From the engineering perspective it is a problem that so **many different domains** are affected:

- Industrial Engineering
- Databases
- Simulation modeling

Industrial engineering skills are in demand to define the requirements for the applications which use the forecast results. Only with background knowledge of the internal behavior of the wafer fab, it is possible to achieve performance improvements. To handle the online data access, a lot of database knowledge is necessary. A major problem is to define the criteria regarding the data quality. From the simulation modeling perspective, the problem is to define the requirements and the level of detail for the simulation model.

Another problem is that all **system requirements are closely related (**Figure 3). The operational applications define the requirements of the whole system. The system requirements define individual requirements of each functional module.

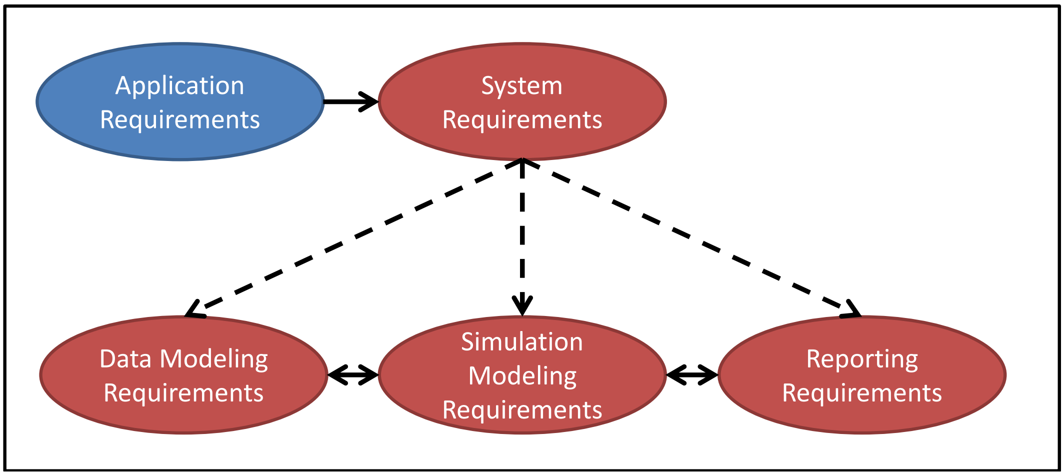


Figure 3: System requirements overview

The reporting module as a part of the whole system is a direct interface to the application. So the applications define the required forecast parameter. The simulation model generates these parameters. Therefore the requirements for the reporting parameters define the simulation modeling requirements. The simulation modeling requirements define the data requirements. The definition of requirements has the opposite direction of the data flow.

2.1.2 Concepts

The problem of the conceptual design is that so many different requirements exist. A solid online simulation **concept needs to satisfy all of these requirements**. Most requirements affect all functional modules. Some requirements contradict each other, for example a high level of detail and the real time requirements. Both requirements are important and it is necessary to find a good balance for online simulation. The conceptual design need to address the following requirements:

- Real time requirements
- High level of detail
- High accuracy
- Maintainability

The **real time** requirement is essential for online simulation. The results for a short time horizon need to be available before they are outdated. Applications that use forecast results have a certain point in time when a decision is required. To satisfy the real time requirement a fast data handling and a fast execution of the simulation model is necessary. All of these components need to be highly integrated.

A **high level of detail** is required because the results for online simulation are used for many small work units in the fab. The production area that uses online simulation is not interested in fab results. The ranges of interest are performance results for a single work center. It is also hard to define a proper level of detail for online simulations. The level of detail for operational applications is ranged between execution level and planning level. So for modeling it is hard to derive the best level of detail because of the high diversity of requirements. For resources, the level of detail ranges from equipment component level, via equipment and work center level, up to the fab level. The desired level of detail for flow items ranges from lot level up to product level. The required time horizon is only a few hours, days up to month.

Another requirement is to reach a **high accuracy** of the forecast results. A prediction needs to reflect the future performance behavior. When a forecast has a large forecast error, it is not useful. To reach high model accuracy, a proper initialization of the simulation model with the most current fab data is necessary. In the model it is necessary to handle disturbances and variability of the wafer fab. To reach a high accuracy a proper **data quality** is relevant too. The model results are only as good as the underlying data.

The **maintenance** aspect is a critical requirement for online simulation. The conditions in the fab are always changing. These changes affect online simulation results. It is necessary to handle those changes. Problems need to be fixed fast, to keep the simulation accuracy on a high level.

2.1.3 Implementation

For the implementation of an online simulation system several problems exist. The key problem for the overall system is the large size and complexity. Numerous of elements are part of the data model. The size of the simulation model gets very large. Many data and simulation modeling elements exist, like lots, routes, and work centers. Many system components exist. The complexity is high, because so many system components and model elements interact. A single change in one system component is capable to have a significant effect on the modeling results. For this thesis data modeling and model validation are crucial parts. The related problems will be discussed in detail.

The major problem for the implementation of the simulation model is the model validation part. The objective is to reach accurate simulation forecast results. The large size and complexity makes the model validation process time consuming. The model size is a problem because a large model contains many error reasons that need to be solved. The complexity is a problem, because the model validation process, especially the identification of the error reasons, becomes time consuming. For many errors a distance exist between the error reason and the effect of an error. If one work center has problems, the forecast quality of another work center is affected much. Secondly, multiple reasons for the deviation from reality need to be checked, see Figure 4. For example for one work center it is necessary to check if the underlying data are all correct, if a stochastic effect like an equipment down causes the deviation, or if a modeling error exists.

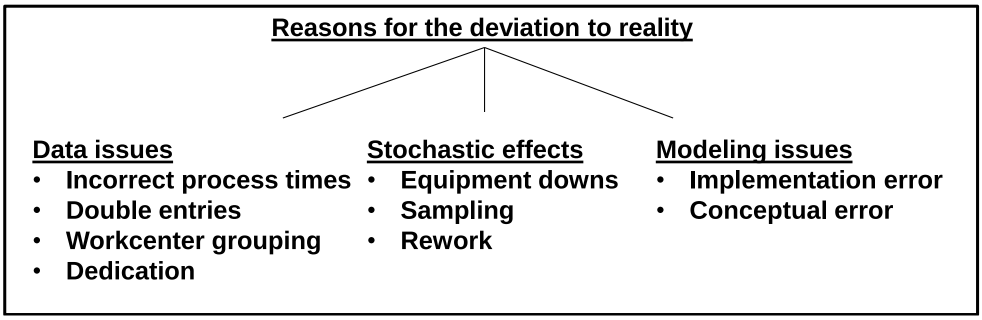


Figure 4: Example reasons for deviation to reality of the forecast results

Regarding to the data modeling part, the problem is that the data is distributed to many databases. Databases contain inconsistent information. Data is often not available in the desired level of detail. In several cases the data is incorrect, missing, or incomplete. For other cases the data volume is very large. These problems have to be solved to create a valid simulation model from the fab data. For online simulation Randell and Bolmsjö (2001) pointed out that beside the model validation, the most effort is spend on the data integration process. In automotive industry the effort for data modeling is also massive (Jensen 2007). The data integration process is highly important to support the simulation model with sufficient data and in a sufficient quality.

Especially when multiple data sources have to be combined the data integration process is crucial. Numerous data integration conflicts occur and need to be resolved. Data integration conflicts are distinguished between schematic problems and data value conflicts (Rahm and Do 2000). Data value conflicts refer to the table values. Conflicts are typing errors, redundancy, contradiction, incorrect data, duplicates, missing values, and problems with different timings. Schema conflicts occur when different data structures represent similar information. These conflicts appear between two relations, between two attributes and between tables and attributes. Conflicts are distinguished between single source and multisource problems. For single sources, quality issues as mentioned above arise within one database source. Multisource data problems and multisource schema problems occur between several different schemas. Even if each individual schema is free from problems, the combination of multiple schemas may cause data integration problems like inconsistency or missing values. The reason for inconsistency is that the data basis was made to satisfy specific needs. Within these categories a wide range of conflicts exist (Table 1). Common examples are naming conflicts like synonyms or homonyms for table names, attribute names and data values. A synonym describes the same object with a different word. A homonym is one word with the same meaning for two different objects. In schema matching literature mentioned above even more detailed categories and conflicts exist.

Data value conflicts	**Schema conflicts**
Naming conflict Duplicates Correctness Missing values Redundancy Contradictions	Naming conflict Formatting Integrity Uniqueness Accuracy/Precision Redundancy Overlapping Default values Unit conflicts

Table 1: Data conflict categories

2.1.4 Validation

The last problem is to do a proper model validation. For the data model itself, the problem is to guarantee a high quality of data. For the simulation model, the major task is to reduce the forecast error between simulation and reality. The problem is to analyze the reasons for the deviation between forecast and reality. Once it is clear what the error reasons are, it is necessary to take actions to reduce them. The problem is to implement an efficient process to execute this model validation task. Due to the high complexity and the large size of the wafer fab, numerous problems need to be fixed.

2.2 Simulation Modeling Accuracy

The objective is to maximize the forecast accuracy. For online simulation in semiconductor manufacturing multiple problems exist which reduce the forecast accuracy. Beside the data problems, three major problems are part of this thesis:

- Warm-up period after simulation start
- Defining the level of detail for modeling features
- Stochastic effects

Model initialization

One big problem for the modeling accuracy is the warm-up period. For steady state models an easy solution is to cut of the warm-up period (Robinson 2004). If the model is not initialized properly, then the time period, until the steady state is reached, will not be used for the results collection. For online simulation this solution is not applicable. Online simulation for short term forecasting has a simulation horizon of only a few days. The transient model behavior is in the range of interest and not the steady state (Reijers and Aalst 1999).

For online simulation the problem is to identify the reasons for the deviation from reality. It is necessary to figure out how big the deviation from reality is, even for well initialized models. Only if these reasons have been analyzed it is possible to further increase the initialization quality and reduce the negative effects on modeling accuracy. So a related problem is to find out, to what extend it is necessary to increase the level of detail for model initiation.

Modeling features and the level of detail

A simulation model for a wafer fab contains numerous modeling features. The problem is to identify which elements of the wafer fab need to be a part of the model and which do not. So the problem is to decide which modeling features have high influence on modeling accuracy and which do not. It is necessary to define a proper level of modeling detail for those elements (Law and Kelton 1999).

Much effort is required to implement features which highly affect the accuracy of the simulation model. The level of detail of such features is high to maximize the model accuracy. Simulation model feature with a low effect on accuracy will be implemented in an abstract way.

An example of an important modeling feature is work center modeling. The work center model highly effects the process time and queue time of the lots. By having proper work center models, the deviation between forecast and real world results will be reduced. Multiple other modeling requirements are necessary as well. Examples are reentrant product flows, rework, and sampling. They also characterize the wafer fab.

Stochastic Effects

In a complex wafer fab a high degree of uncertainty exists. Multiple events are unpredictable but still have an influence on the fab performance. Examples are:

- Equipment downs
- Sampling
- Hold
- Rework
- Split
- Variable process times

These events cause a high variability in the wafer fab but also in the simulation model. So the problem is to handle this variability in the model and minimize its effect on the forecast accuracy.

2.3 Real Time Capability

Real time capability means, that the results are available before they are outdated. The real time requirement is a serious problem because the online simulation needs a high detailed, well initialized simulation model of the full fab. The computation time includes the time for:

- Data access to extract the most current fab data
- Data transformation and cleaning
- Model generation
- Simulation run, including the results generation

The computation time is high because data processing and the simulation run itself are time consuming operations.

In the context of online simulation two different levels for the real time requirements exist. These levels depend on the particular way how the forecast is used. The first use case excludes human interaction. The time constraint for the particular application is the limiting factor. The second use case includes direct user interaction. The limiting factor is the human tolerance to wait for the results.

The first use case excludes direct user interaction. An automated scheduler triggers the online simulation forecast periodically. The simulation model will be initialized with the current fab state at this point in time. Figure 5 depicts the timeline for a forecast. The computation time is the time between the forecast trigger and the time when the results are available. The forecast period is the time between the forecast trigger and the maximum forecast horizon. Due to the computation time, the first part of the forecast period already passed in reality. So the results for this time period are obsolete. For time critical applications the danger is that the computation time is higher than the application specific forecast period.

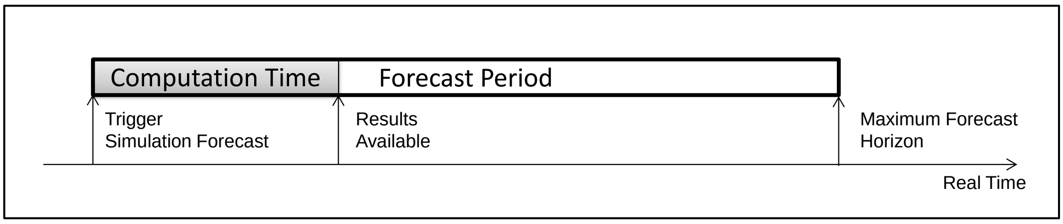


Figure 5: Timeline for Forecasting

The problem is to provide results before they are outdated. The point in time when the data is outdated depends on the particular application. In general an acceptable time limit is a few minutes.

The second use case includes user interaction. For this case, the user triggers the simulation to run. This means that the user initializes a model with the current fab state and executes the simulation. The user is also able the change the model settings beforehand to simulate different scenarios. After triggering the simulation run, the user analyzes the simulation results. It becomes clear that the real time requirements are much higher compared to the first use case. The user has to wait until computation is finished. The problem is to reduce the computation time to an acceptable time limit of only a few seconds. The objective is that the user has to wait no time or only little time for the results. Only if the waiting time reaches that level, this use case is acceptable.

2.4 Online Simulation Applications

A plain simulation forecast without any follow up actions has only little value (Walonick 1993). Therefore several applications are required on the shop floor to use the forecast results. Those applications take actions in the manufacturing process to improve the factory performance.

For semiconductor manufacturing there are many applications available to improve factory performance. The first problem is to identify those applications. The second problem is to check, if online simulation forecast results are useful for them.

The identification of applications is a problem because of the high complexity and large size of the wafer fab. Many approaches exist but the knowledge is spread among several people. The decision process it is often not clearly defined. In daily operations the people execute the decision process automatically. Often a solid data source does not exist. A literature search is also useful. The expectation is that those applications cannot be introduced without modifications. Different wafer fabs also differ in the specific ways they execute their decision processes. Not all applications from literature are feasible in all wafer fabs due to different properties of the manufacturing process itself, a different degree of automation, a different quality of the data sources, different problem areas, and different objectives.

The second problem is to check, if the identified applications are compatible with online simulation. This means, every application has certain requirements. Only when the forecast meets these requirements, it is feasible to implement this application in the wafer fab. The criteria, to measure these requirements, are:

- Computation time to generate results
- Time horizon of the forecast
- Level of detail for the forecast
- Deviation from reality

If it takes too much time, to compute the forecast results, the chance is high that the results are already obsolete. In this case an application is not feasible. If the forecast time is not long enough to implement changes in the fab, then such an application is also not feasible.

Another problem is that these requirements are not clearly defined in most cases. First of all, nobody can exactly define which deviation of forecasting results is still acceptable. Secondly there is no formal business process. The requirements vary for each individual decision with the same decision variable. Due to the high complexity in the wafer fab, the variability of these requirements is also high. For the same decision variable multiple dependencies exist.

3 Related Work

3.1 State-the-Art of Online Simulation

To address the needs of an intelligent operational fab management multiple methods from different research areas are available. This chapter provides a literature overview of most common forecasting, modeling, and optimization approaches. The current state of the art will be presented, including real world examples to solve operational problems. This chapter also distinguishes between common approaches for scheduling, operational control, and planning. Based on the current state of the art, the desired target system will be presented in detail.

3.1.1 General Modeling and Optimization Approaches

In industry and science solution approaches exist with the objective to achieve performance improvements. In operations research literature a wide range of exact mathematical methods is available (Neuman and Morlock 2002). Examples are linear programming, nonlinear and mixed integer programming, branch and bound, or dynamic programming (Denardo 2003). Mathematical formulas describe the model elements, like constraints or objective functions. Pinedo (2002) describes scheduling methods. Many heuristics are also used due to an increasing level of complexity (Pearl 1984). Examples are metaheuristics and genetic algorithms. They combine randomness and prior knowledge of available solutions to find improved solutions.

Many modeling approaches are used in industry and science like Petri nets, Markov chains, and queuing theory (Reisig 1985, Lindemann 1998, Girault and Valk 2002, Norris 1998). Figure 6 is illustrating Petri net, Markov chain and queuing models. For Petri nets, places and tokens represent model states. Transitions make state changes possible. The tokens can change their place. For Markov chains, each number represents a model state. The transition state represents the probability of state changes. For the illustrated queue model, mathematical formulas are used to describe arrival, departure, queue and processing behavior.

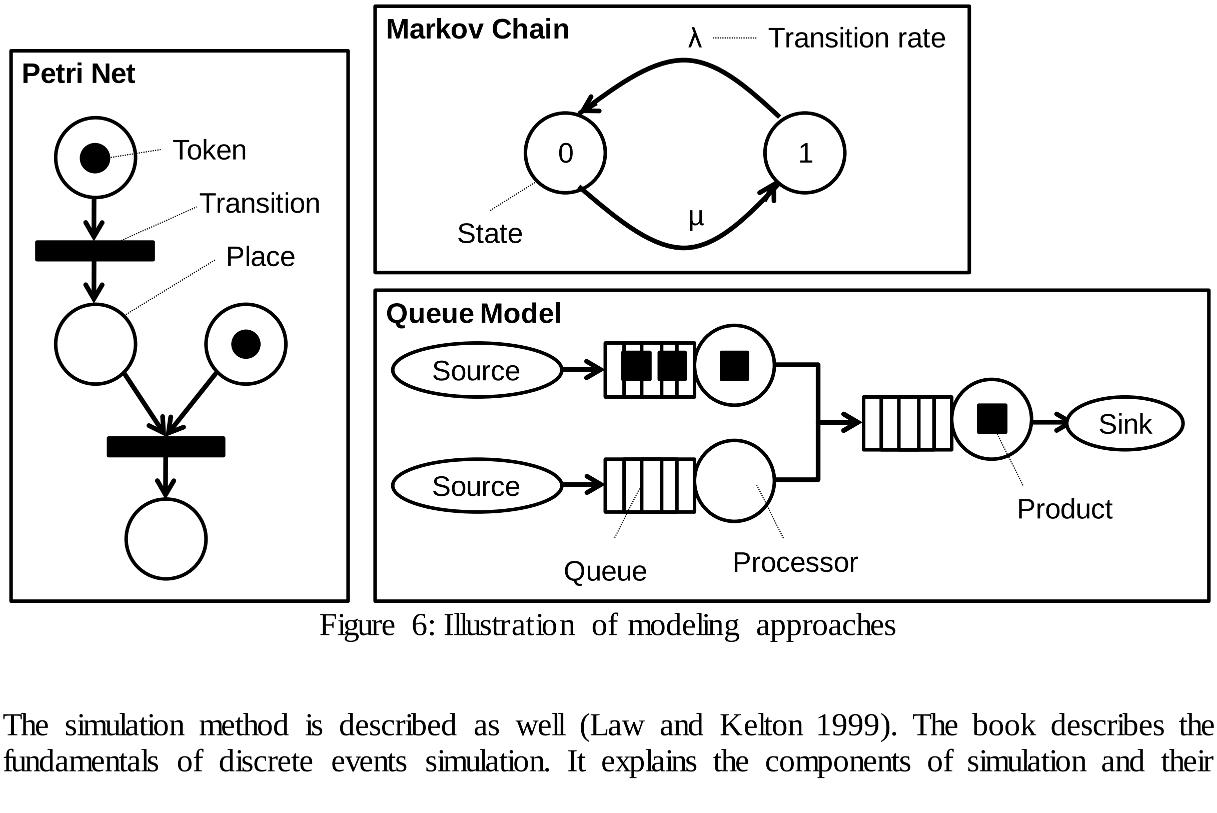


Figure 6: Illustration of modeling approaches

The simulation method is described as well (Law and Kelton 1999). The book describes the fundamentals of discrete events simulation. It explains the components of simulation and their interactions. It also contains methods for experimental design and analysis. The interaction of simulation and optimization is described (März et al. 2011). It provides a good overview how simulation and optimization interact. It also contains several examples for real world applications. The objective of simulation based optimization is to achieve real world performance improvements. The simulation part mimics the real world behavior. The optimization part creates and evaluates multiple different scenarios, with different settings in the simulation model. The settings of the scenario with the best performance will be applied to the real world.

In this section it becomes clear, that a wide variety of general methods exist. Each method has its advantages and disadvantages. The success of a method also highly depends on the background knowledge of the people, who apply it. Besides that it is useful to extend the point of view to the forecast perspective to address operational workflow problems. So in the next section, several literature examples will be presented, which are mainly used for the purpose of forecasting.

3.1.2 General Forecasting Approaches

Pure forecasting methods are employed in different domains for a wide range of applications, for example weather forecast, business forecast, physics, biology etc. In business, a forecast is used to predict for example sales and tax income (Wilson and Keating 1994). For weather forecasting, methods like analog and numerical weather prediction methods exist (Holton 2004). It is obvious that forecasting is highly useful because it is applicable in many different domains.

According to Walonick (1993), a forecast is useful when a change to the predicted process outcomes is intended. So, forecasting becomes interesting when there are ways to modify the forecast itself or the effects of the forecasted elements. These conditions exist for operational decision making. Therefore forecasting becomes interesting for operational fab control. As mentioned before, the task is not only to build a forecast system, but also to find ways to modify and improve the future.

When having a closer look at forecasting methods, it becomes obvious that the preferred methods are also domain specific. Walonick (1993) provides a very good overview of general forecasting methods. Examples are intuition, questionnaire, interview, and consensus. In semiconductor manufacturing these methods are obviously not suitable for daily decision support. Other methods, to figure out future trends, are historical analysis with historical trend extrapolation, Box-Jenkins method, and regressions. Mertens and Rässler (2004) list and describe several mathematical methods which are used for forecasting. Other forecasting methods are also simulation and mathematical models. Their advantage is that they have internal knowledge about the system relationships. They are applicable for math describable worlds, like semiconductor manufacturing.

In the manufacturing context, specific forecast approaches are available (Aitchison and Dunsmore 1975, Chih-Yuan Yu and Han-Pang Huang 2002, Mosinski et al. 2011). The general concept is to have knowledge about historical fab performance. In addition to that, it is necessary to capture the current parameters, which have the most influence on that forecast. The basic concept is depicted in Figure 7. To illustrate the concept, a simple example will help. The task is to predict the remaining cycle time of a lot until it completes production. This represents the output parameter y. The parameters, which have most influence on the prediction, are the current lot position (x1) and the lot priority (x2). Historical knowledge is available about the average cycle time of a lot, from its current operation until it finishes production (c1). The influence of the lot priority (c2) is also available in history. With the historical information and the current lot information, the remaining cycle time of the lot (y) will be computed. This is a very basic forecast principle, commonly used in practice.

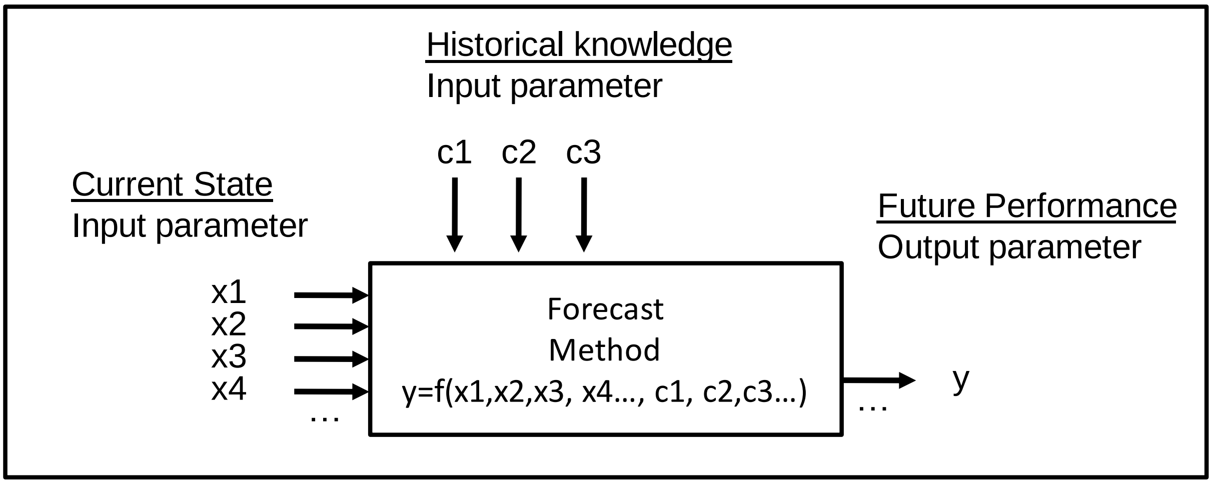


Figure 7: Forecast principle

It is clear, that the general modeling and optimization approaches from the previous section and the general forecasting methods from the current section are established within their application areas. It is useful to have a closer look at specific literature for operational problem solving in semiconductor manufacturing. The target is to find out about the current state of the art. It is useful to figure out which methods exist in literature, to solve specific operational problems.

3.1.3 Approaches for Operational Control of Manufacturing Systems

In Bagchi et al. (2008) a discrete event simulation for fab performance forecast and fab optimization is generated. High forecast accuracy and a low simulation computation time have been achieved. The proposed system is highly useful in the production system. A customized discrete event simulator has been created. The basic data relies on historical data and current fab data as well.

Zisgen (2007) presents a flow modeling approach for short term planning. The advantage is that the method is quite fast because of the high level of abstraction. Compared to discrete event simulation, the flow model approach leads to a high event reduction which needs to be processed. Disadvantage is the information loss due to high abstraction level. Furthermore the discrete world is hard to transfer into a continuous model. The interface for the fab data sources has to be very powerful.

Reijers and Aalst (1999) present a very early approach for short term simulation. The approach has been used for a social security company and not for semiconductor manufacturing. This publication provides a very detailed insight from a theoretical and a practical point of view.

Aalst (1998) describes the use of Petri nets for workflow management. He describes the basic concept of Petri nets and how they are applied in manufacturing. Several ways to for analysis and verification of Petri nets are described as well.

Simulation based scheduling does also combine the advantages of two methods. The objective is to generate an optimized job sequence for the production system (Chong and Sivakumar

2003, Dangelmaier et al. 2006). The sources show that the schedule horizon is only one shift. Rescheduling also becomes important, if disturbances occur.

Chih-Yuan Yu and Han-Pang Huang (2002) present a forecast method using neural networks. Historical knowledge is used to predict future fab performance. Based on multiple factors of the lot like priority, work center queue, and product group a factor for future wait time and process time will be derived. By cumulating these factors more statistic data like work center moves, lot cycle time and product cycle time can be derived.

A very good conceptual overview of using simulation of operational control is available in Drake and Smith (1996). They summarize the challenges and potential benefits of online simulation. Their work focuses on the basic concepts of online simulation. As a result they present a framework that simplifies the creation of online simulation models.

Smith et al. (1994) propose a way to connect simulation with shop floor decisions. By using the same control logic for simulation and shop floor, the implementation effort is reduced. The simulation model is used for scheduling decisions as well as for short term performance prediction.

3.1.4 Method Selection

The process to find operational problems and the corresponding solutions is a bi- directional approach. On one side multiple operational problems exist. On the other side multiple solution approaches are available today. For each operational problem, none or multiple solutions apply. The solution approaches are applicable to none or multiple operational problems.

The objective is to select a method to maximize the benefit for the wafer fab. To maximize the benefit, a method has to increase the fab performance and decrease production cost. To do so, it is necessary to gather as many operational problems as possible because they have a massive impact on the fab performance.

General Approach

At this stage multiple ways are available to address operational problems (Figure 8). One way is to do pure forecasting. Another way is to implement a system to compare different settings, so called “what- if Scenarios”. The third method is optimization, which automatically finds the best settings for certain variables.

The forecasting method requires one model which captures the future fab performance best. The user is able to change certain fab settings and improve fab performance. The decisions are based on the forecast results and the user experience.

A generation of “what-if scenarios” requires a model and a decision parameter, to apply different settings. Based on experience, the user changes the value of the decision parameter, to generate and evaluate different scenarios. Then the user selects the scenario with the best performance and applies the settings in the fab.

An optimization approach needs a model, a decision parameter, an objective, and an intelligent algorithm to find the best solution in the search space. The manual effort to select a scenario and to evaluate it will be reduced. So, modeling, scenario generation and optimization are consecutively built up on each other.

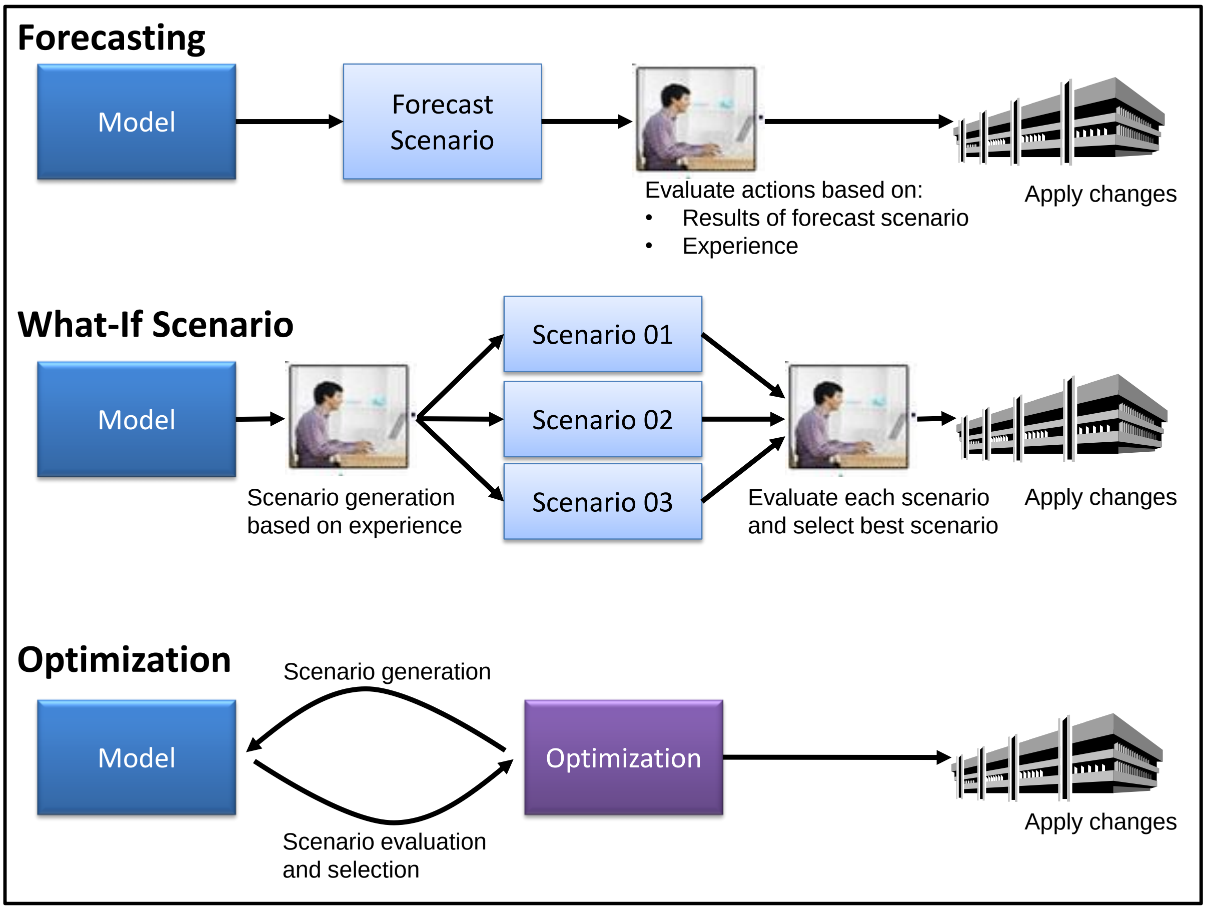


Figure 8: Forecasting, What-If Scenario, and Optimization

Out of practical reasons it has been decided to start with a pure forecasting system. The main reason for that is to gain user acceptance as early as possible in the project phase. A pure optimization system is not very transparent to the user. The user is not deeply involved in the decision making process. There is a risk that the user rejects the system. Especially in the beginning it is necessary to obtain as much user feedback as possible to improve the model. If the user is involved in the development and improvement process of the model, it is easier to receive user acceptance. Once having an expert model for forecasting, several extensions like what-if scenarios and optimization approaches can be added at a later stage. Another reason is that a pure forecast is applicable in many different ways. It is almost impossible to list all improvement areas. The reason is that it requires deep expert knowledge in multiple complex manufacturing areas. What-if scenarios and optimization also require formalizing all the changing parameters. Expert creativity is much more flexible and less limited. All these reasons show the practical advantages of a forecast used as a decision support system.

Comparison of Methods in Literature

Every method has advantages and disadvantages. The success depends on multiple factors like the specific application, the fab environment, the used method implementation, and the personnel's skills. The following literature is a basis for a method selection process.

Klemmt et al. (2008) compares mixed integer programming and simulation based optimization using genetic algorithms. The outcome is that for a small problem size, the mixed integer approach reaches the optimum or is very close. For a large problem size, the simulation based optimization approaches deliver better results. For very large problems the mixed integer approach is not able to find a valid solution. Schuster (2003) shows that large scale problems with a lot of equipment are hard to solve, due to an increasing level of complexity.

Heegaard and Trivedi (2009) compare Markov chain models, Petri net models, and simulation modeling. They have pointed out that analytical solutions become inefficient with an increase in model size. On the other hand they mentioned that simulations become inefficient when the number of events increases.

He, Fu and Marcus (2000) apply a Markov decision process for fab level decision making. It is an extension of the Markov chains approach. It is mentioned, that the danger of a state explosion for a full scale wafer fab exist. Decomposition approaches have been applied to solve the problem.

Robinson (1998) pointed out that their justification of fluid models or diffusion models lies in the heavy traffic theory. They are highly relevant for heavily loaded systems.

Dangelmaier et al. (2006) is dealing with scheduling solutions. He emphasized: "when the frequency of disturbances is high, then management efforts could be better spent on reducing those uncertainties, than developing complicated scheduling logic". For a semiconductor logic fab with high variance this means, that the horizon where scheduling is applicable is rather short.

Pure forecast methods are available to predict future fab performance (Chih-Yuan Yu and Han-Pang Huang 2002, Mosinski et al. 2011). The advantage is that it takes significantly less effort to create a forecast, based on historical trace data compared to carrying out a simulation model. A forecast based on historical data does not need to have deep system knowledge as a simulation run. The disadvantage is that these forecasting methods often lack the internal knowledge of the fab behavior. So these methods are often not sensitive to certain fab characteristics and their relations. One example is the missing consideration of the work center capacity limitation. It decreases the forecast accuracy. Another disadvantage is that the forecast method is fixed to one specific output parameter. By changing the forecasted parameter (like lot cycle time to work center WIP) the whole method needs to be changed. Compared to simulation there is less flexibility to use the results. Furthermore the pure method is hardly extendable, when customized solutions are required, which needs internal knowledge about system relationships. Thus, this method is not sensitive or extendable to certain relations. It is not possible to apply optimization or to run what- if- scenarios.

Decision

The decision is to use simulation because this method matches the problems requirements best. According to the literature above, the biggest advantage of simulation is that it is scalable. Even large problems in term of fab size and simulation horizon are applicable. Simulation and simulation based optimization are capable to deliver reliable results even for large problem sizes. Exact methods like linear or dynamic programming are hardly scalable. Complexity and the search space increase dramatically. To deal with complexity, those exact methods are using decomposition. The additional effort to deal with decomposition is not necessary in simulation. For short term simulation, the concern of event explosion, as mentioned for heavily loaded systems, does not apply, even for large problem sizes. The reason is that a simulation model for fab operations will only simulate a few days. It will not simulate several months or years like a simulation model for planning purposes.

Additionally, the interface for online fab data in simulation is intuitive. In discrete event simulation the model entities like equipment or lots are the same as in the real world. Compared to math or fluid models, a complex parameter transformation is not necessary. When there is no complex transformation step required, the implementation of the fab interface becomes fast. The model validation becomes more intuitive, when comparing results to the real fab.

For simulation, the level of detail is very flexible because a simulation model on lot level and equipment level has many details. It is usable for attaining detailed results but also for very abstract results. The key is to apply different abstraction levels to the reporting. For fluid models, deep changes are required to increase the level of detail, once the model has been created.

Another advantage is that simulation models are flexible to run for standalone forecast or to be applied in what-if analysis and optimization later on. Discrete event simulation is also extendable to certain needs, if it becomes necessary.

In order to address the full scope of operational solutions Kohn et al. (2009) describe a mix of several approaches. The objective is to address different needs of operational applications.

3.1.5 Automated Model Generation

Beside modeling and forecast methods, the capability of an automated model generation is one component of online simulation. For online simulation, the model generation part is a time critical and repetitive process. It is obvious, that it is not feasible to execute it manually.

Mathewson (1984) provides a definition for automated model generation. He defines that it is software that translates the logic of a model into the code of a simulation model, whereby a computer is able to reflect the model behavior.

The first idea of automated model generation came up very early (Oldfather et al. 1969). A way has been proposed, to generate computer programs out of a questionnaire.

In the context of shop floor control, also Son and Wysk (2001) elaborate the aspect of automated simulation model generation. They provide a method, to generate simulation models out of the resource model and out of the control model of the shop floor. They also provide several examples to illustrate and validate their methodology.

Automated model generation is up to date in science and industry. Buzzwords like “digital fab” dominate future challenges to increase productivity and enable a high degree of automation. Automated model generation is one part of it. Today different manufacturing areas, like automotive industry or shipyard manufacturing apply automated model generation, to enable interaction of fab data and model data (Jensen 2007, Burnett 2008).

For this thesis, the implementation of automated model generation will not be elaborated further. A variety of sources in literature is already available, with the earliest originating from 1969.
Within the context of this thesis, the core of automated model generation is the implementation of an interface, to export the data from one format, the database schema, to another format, the simulation model files. Chapter 4 and 5 describe the data modeling aspects and the simulation model aspects in detail. The implementation of the transformation step itself is not part of this thesis.

3.2 Data Modeling for Online Simulation

Data modeling is extremely important for online simulation. McNally and Heavey (2004), but also Robertson and Perera (2002) pointed out that the quality of the model results closely depends on the input data. For online simulation several data modeling aspects are in the range of interest:

- Required data content
- Real time requirements
- Data integration
- Data quality

The following literature addresses these aspects. The first part provides an overview of common database literature. Here are the data integration aspects in the range of interest. The second part contains literature about data handling in the context of modeling and simulation.

In database literature there are many sources available about data integration, data warehousing (DWH), and extract transform load (ETL). A classification of data quality problems with multiple examples is available in literature (Rahm and Do 2000). Data cleaning approaches are presented. Lehner (2003) provides an overview of data warehousing approaches. Data integration and data consistency for data warehousing is discussed in detail. Rahm and Bernstein (2001) and Wang and Murphy (2004) provide an overview of automatic schema matching approaches. A classification of schema matching approaches is available. Multiple approaches from literature will be characterized.

For online simulation it is further necessary, to deal with more specific data integration literature for simulation purposes. Randell and Bolmsjö (2001) pointed out the requirements of databases for online simulation. They highlight the aspect of high speed, usability maintenance, scalability, software independence and less manual intervention for simulation. Skoogh and Johansson (2008) are dealing with data input management for discrete event simulation. The major focus of this work is the data quality, especially the right level of detail for the data and the data collection. The whole process to provide data for simulation models and improve the data quality is presented in detail.
Additional work has been done to reduce the computation time and effort to provide valid data for simulation modeling (Horn 2008, Horn et al. 2005). He analyzes the input data management from a semiconductor perspective. For Skoogh et al. (2010) the manufacturing environment is the automotive industry and the aerospace industry. He presents ways to create simulation models from manufacturing data.
Furthermore Skoogh (2011) describes the current state of the art of input data management for discrete event simulation in detail. He has pointed out that the input data related activities consume more than 30% of the time in simulation projects. Another fact is that the degree of automation is rising but still very low. About 20% of the companies which are using discrete event simulation have automated solutions for the input data management.

3.3 Literature for Accuracy

To reach a high accuracy for online simulation for semiconductor manufacturing, it is first necessary to model the fundamental elements of the wafer fab (Atherton and Atherton 1995). Law and Kelton (1999) have also pointed out basic modeling features, like route, product, and equipment modeling. One way to increase accuracy is a more precise equipment modeling. Dümmler (2004) describes the complexity of a cluster tool and modeling approaches on a high level of detail. Another element used to increase the accuracy is the dispatching rule modeling (Sivakumar and Chong 2001). An accurate dispatching rule modeling is highly important to model the fab behavior. Altogether it becomes clear, that one way to increase the modeling accuracy is to increase the accuracy of each simulation feature. For online simulation in semiconductor manufacturing Bagchi et al. (2008) summarize the important model components. Common modeling components are for example, dispatching rules, equipment tool types, setup, equipment downs, hold, and sampling. He also emphasizes the importance of an accurate model initialization, including the majority of the WIP lots, hold states, and equipment states.

Besides the modeling content the validation process is also one task to be executed for increasing the modeling accuracy. Several model validation methods are available (Rabe et al 2007, Balci 1998). It is clear that for online simulation the model validation aspects are important to increase the accuracy (Bagchi et al. 2008). The online simulation model has been validated by validating equipment throughput, fab throughput, flow factors and lot traces against real data.

Another aspect of modeling accuracy is the model initialization. The general problem is that the simulation results right after simulation start (warm-up period) do not reflect the behavior of the real system. In science a lot of literature is available about the warm-up period (Hoad et al. 2008). It turns out, that most literature sources are looking for better ways to estimate the length of the warm-up period (Wilson and Pritsker 1978, Schruben et al. 1983, Robinson 2002, Mahajan and Ingalls 2004). Many methods exist to deal with incorrect results during the warm-up period (Robinson 2004):

- Cut off the warm-up period
- Initialize the model very well
- Combination of model initialization and cut off the warm-up period
- Increase run length to reduce the proportion of the warm-up period
- Estimate the steady state from a short transient run

It is necessary to initialize the model with the current fab state in order to increase the accuracy for online simulation. In literature Reijers and Aalst (1999) have explained the high importance of the model initialization. The accuracy of simulation for short time horizon highly depends on the initial state. So the focus of this thesis is to analyze the accuracy of simulation results during the warm-up period, even if the model has been initialized very well.

3.4 Literature for Applications

In semiconductor manufacturing discrete event simulation is quite established to address several applications, especially planning problems. Rose (2006 b) optimized the overall fab performance by changing the equipment capacity, the dispatching rules, and the lot release by using simulation based optimization. Scholl (2008) analyzed disturbances like changing product mix, storage effect and breakdowns. Rose (2006 a) analyzed effects of an intermediate lot storage in a semiconductor wafer fab. Sivakumar and Chong (2001) analyzed the effect of lot size and dispatching policy. Klein et al. (2006) considered product ramp up effects in manufacturing. Dümmler (1999), Niedermayer and Rose (2003) determined and improved cluster tool cycle time. Roser et al. (2001) applied simulation to verify a method for bottleneck detection. Domaschke et al. (1998) worked with route changes, batch policy changes, dedication changes, lot release changes, and transport changes to increase the fab performance.

For planning purposes a wide range of literature is available about simulation applications and performance improvements. Operational problems, like dispatch rule selection, are solved by offline models. Zhang et al. (2008) propose a way to use a dynamic dispatch rule selection by using steady state simulation to create a response surface. The response surface methodology is able to reflect different fab conditions. It is used to validate and configure the dynamic dispatch rule selection algorithm. For different fab conditions, different dispatch rules apply. Therefore offline steady state models are also useful for operational problem solving.

In literature several sources are available to address manufacturing applications in a dynamic manner, using online simulation. A simulation based dynamic dispatch rule selection solution is available (Wu and Wysk 1989). The objective is to apply the dispatching rule, which performs best to the current fab state. Simulation based scheduling approaches are available (Horn et al. 2006, Potoradi et al. 2002, Klemmt et al. 2008). The purpose of this kind of scheduling is to increase throughput and to reduce cycle time. März et al. (2011) present a strategy to couple simulation and optimization. The purpose of the simulation results is to provide good starting values for optimization. Roser et al. (2001) present an approach to identify bottlenecks in the manufacturing area. The basic idea behind this is to find the bottlenecks first, before it is possible to improve whole fab.

For the automotive industry Müller Sommer (2012) describes different applications for planning purposes and for operational purposes. For planning purposes, typical applications are layout planning and capacity planning. The objective is to find the optimal fab configuration for resources, processes, and products. For operational purposes the improvement of existing systems is the main objective (Jensen 2007). Typical applications are setup reduction, work center reconfiguration and operator assignment. For the operator assignment it is a key challenge to deal with different qualifications. Additional applications determine the optimal storage size and lot size.

To increase the performance improvement for online simulations it is necessary to obtain the full picture for a wide range of operational applications. So for this thesis a systematic analysis of several applications will be addressed. The main focus is to analyze the capability of an online simulation forecast to enable these applications.

4 Data Input Modeling

This chapter addresses the data input modeling of an online simulation system. It contains the data requirements and the data model design decisions. The created data schema will be presented. Data integration problems and related solutions are also available. A summary of the data modeling aspects for short term simulation has been published in (Noack et al. 2010).

Data modeling is highly important for online simulation. The simulation model will be generated from this data. Therefore the simulation results are only as good as the underlying data (Robertson and Perera 2002). An automated data processing has two major advantages compared to a manual data processing. Firstly an automated solution is very fast to provide the most current fab data for the simulation model. The real time requirement is essential for online simulation. Secondly the creation of a forecast and the data processing are repetitive tasks. An automated solution is suitable to process the data again and again without much effort.

4.1 Data Requirements

In terms of data requirements multiple aspects like data content, data quality and data integration are important. First of all, it is necessary to define which data is required. The data content requirements highly depend on the simulation modeling requirement. Secondly a proper quality definition for the input data is also required. Without a data quality definition it is not possible to distinguish, if it is feasible to create an online simulation model out of the underlying fab data. It is also required to define how the data input will be integrated in the whole simulation system (Figure 9). The data model is one component of the whole forecasting system. It connects the simulation model to the underlying fab data. Multiple design decisions are necessary to define, to connect, to transfer, and to transform the data.

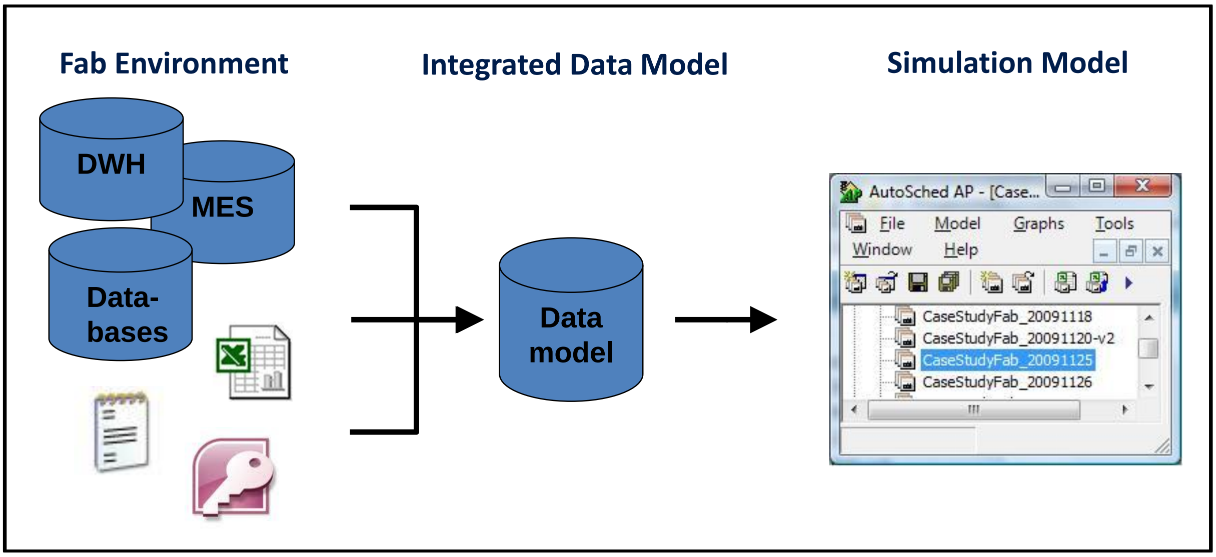


Figure 9: The data model as a connection between fab data sources and simulation model

4.1.1 Data System Requirements

The most important requirement for the data system is to provide the simulation model with the necessary **data content** and **data quality**. The next two sections will further elaborate on these requirements. From a system perspective the following aspects are very important.

The major requirement of the data system is to provide data at **high speed**. A high speed is critical to achieve real time analysis capability. It is necessary to execute the data extraction process, the data transformation steps, and the loading process (ETL) very fast.

A high degree of **automation** is also required. The simulation forecast is frequently demanded. It is not feasible to transfer and transform the data manually from the data source into the simulation model. The labor costs will get too high to run a short term simulation. This repetitive task has to be automated. A manual interference will be reduced to a minimum.

Another requirement is a high **availability**. This means, that the data model and the data sources for online simulation have to be available at any time. This is a challenge because online simulation needs a high diversity of fab data. If the underlying data sources are not available, a forecast is not possible.

The requirement of **reproducibility** makes it possible to generate historical models and simulation results again. It is highly recommended for model validation and error checking. In a changing fab environment, the data content changes as well. Several data sources do not capture historical fab states. They contain the most current state only. In some cases, when an error in a simulation run has been identified, it is necessary to figure out where the error comes from. It is possible that the origin of the error is the simulation model, the data model or the fab data sources. If the underlying data in the data model has been changed already, it is not possible to trace an error back to its origin.

When reaching the roll-out stage, the **maintenance** aspect becomes important. Model updates, new requirements, data source schema changes, and changing links to data sources have to be applied. The data quality needs to remain on a high level. The quality of the simulation forecast results closely depends on the data quality. A proper documentation is necessary as well. It is very valuable to get an overview of the data sources, the related transformation steps, and the target data structure.

4.1.2 Data Quality Requirements

A data quality aspect for online simulation is to have **complete data**. If data is incomplete or entirely missing, the simulation behavior becomes unpredictable. There is a risk, that the simulation run does not finish successfully, or that the forecast error is large.

The **data accuracy** is another issue of consideration. It directly affects the accuracy of the simulation results. If the data values are inaccurate and they do not reflect the reality, the simulation results are also not correct.

Another aspect in terms of data quality is **data integrity**. Data also should not contradict each other. An even more detailed aspect is to reach referential integrity. Valid relationships are important for data in different data tables.

For online simulation it is important to obtain the most current fab state because the transient simulation approach highly depends on the initial state. Therefore the used data values have to be highly **up-to-date**. In industrial environments ongoing changes occur everywhere. Typically the changes can be classified into two major categories: dynamic and static changes. Examples of dynamic changes are the equipment states, the current lot position, the current PM plan, and the current lot release plan. These data values are supposed to change frequently. Static changes are the route definitions, the equipment process times, and the dedication matrices. They are not changing very often. Moreover planning information about the future, like the lot release plans and the PM plans are limited to a time horizon.

The **level of detail** of the data is also important. Finally it is required that the level of detail of the data is matching the level of detail in simulation. When the level of detail is too high, the complexity increases. When it is too low, the accuracy of simulation results is reduced. To match the simulation modeling requirements the data has to be on the same level of detail or a data transformation is required.

A criterion worth striving for is the avoidance of **unnecessary data**. Data that is not relevant for modeling does not need to be a part of the data model. Such values increase the data volume, and decrease the transfer speed. Those values will often lead to inconsistencies.

4.1.3 Data Content Requirements

This section describes the data requirements for online simulation in a semiconductor manufacturing environment. Each simulation modeling feature needs a certain amount of fab data. Most of this required data is also quite close to common steady state simulation models. The following tables summarize the data requirements for short term simulation. The next section elaborates on the different data requirements of a steady state simulation for a planning purpose and a short term simulation for operational decision making.

Table 2 contains lot related information. Lots are the flow items in the wafer fab. They contain the wafers, which are processed in the wafer fab. Information about current WIP lots and future lot releases are necessary.

Property	Description
Lot name	The lot is a box which contains several wafers. The lot name is required to identify one lot. The lot name needs to be unique.
Quantity	The quantity defines the number of wafers in a lot. It is necessary to handle wafer depended statistics and wafer dependent process time.
Route	The route contains the sequence of process steps to create the final product. Each process step requires resources. Every lot is assigned to one route. Every lot follows this step sequence.
Product	The product is a grouping concept to define which lots have similar properties. Lots with similar properties (for example routes or process times) are in the same product group, lots with different properties are in different product groups. Product information, with different level of detail, enables additional grouping concepts for lots.
Lot priority	The lot priority is highly important for dispatching. Lots with higher priority will be processed first.
Release time	The release time is the point in time, when the lot enters the wafer fab. It is necessary to create cycle time statistics.
Due date	The due date is the time target when the lot has to finish production in the wafer fab. The fab due date is required for dispatching rules with time oriented targets.
Current step	The current step defines the progress of the lot within its step sequence at a particular time. Every WIP lot in a fab has a current operation. It is essential to initialize the fab.
Current equipment	If the lot is processing, the equipment name identifies the resource of the lot. This field is empty if the lot is not processing.
Remaining process time	If the lot is processing the remaining process time indicate when the lot will finish the current process.
Lot attributes	Several attributes are required to address specific behavior of this lot. Lot attributes contains information about dedication, sampling, etc.
Current lot exceptions	Lot exceptions, like hold, rework, or split are required for lot initialization. With this additional exception information, the point in time where the lot will follow its regular flow is determined more precisely.

Table 2: Lot data requirements

The route information is required to model the process flow of the lots. Route related data requirement are available in Table 3.

Property	Description
Route name	The name of the route identifies the step sequence. The route name needs to be unique.
Operation	The name of a single process step within the step sequence.
Required resources	The main require resource in the wafer fab is the equipment. Examples are lithography, wet bench, or implantation equipment.
Additional resources	Other resources, beside equipment, are also used to execute a process step. A limited resource is photo mask (reticle). It is necessary to execute a lithography step.
Sampling	Sampling is used to trigger optional process steps. It is often used for measurement. Not every lot needs a measurement to extract information about process performance. The rate determines the probability whether a lot will execute a particular step.
Split and merge	Split and merge tasks separate or combine a lot, whereby a new lot will be created or an existing lot will be deleted. A lot will execute split or merge, if some wafers of the lot execute different steps than the other wafers. Further information about the split and combine operation, the probability and the number of wafer in a lot are required.
Hold	Hold defines that the lot will stop processing. The lot does not proceed with its regular flow.
Rework	Rework is necessary, if a process step fails. It needs to be done again. Further information about rework probability, rework route, and return operation is required.
Transportation	The transportation information is required to model the cycle time delay of the lot, when the lot is changing its position to get to the next equipment.
Operation target duration	This is the planned duration for an operation. The duration is used for time target oriented dispatch rules like ODD.
Time window constraints	Due to process constraints, the lots are forced to execute several operations within a limited time. Information is required to model such process constraints.
Additional constraints	Several constraints affect the process flow of a lot. Depending on the particular simulation modeling requirement such information is necessary too.

Table 3: Route data requirements

Equipment is the resources of the wafer fab. The process flow defines which steps require which resources. Table 4 contains an overview of equipment related data requirement.

Property	Description
Equipment name	An equipment name is required to identify a resource. The equipment name needs to be unique.
Equipment group	An equipment grouping concept is required to create statistics of several equipment. It is also useful for resource allocation, whereby a process step is assigned to an equipment group, which is able to handle the process.
Dedication	Dedication is a concept whereby not all equipment in the same equipment group is capable to execute every process. This is the case, even if the hardware of the equipment is the same.
Setup	Setup defines the required equipment state for a process step. If the required setup state needs to be changed, it causes idle times for the equipment. To model setup, additional information is necessary, for example the current setup state of the equipment, the required setup state for the next lot, and the time to change the setup for the equipment.
Process time	The process time characterizes the time of a lot executing an action in the related equipment. Process time is defined per wafer, per lot, or per batch. It depends on many factors like the number of wafer per lot, the product number of the lot, or the equipment.
Throughput	The throughput of the equipment provides information about how many wafers, lots, or batches can be processed within a time period.
Equipment model characteristics	Equipment characteristics provide more detailed information to model the equipment behavior with more details. Several factors affect the equipment modeling: • Maximum number of processes or batches • Maximum and minimum batch size, to define how many lots enter one process at the same time. • Number of processing chambers
Dispatch rules	Dispatching rules prioritize the lots and assign them to equipment. The dispatching decision depends on many factors like lot priority, setup constraints, or batch size constraints.
Unscheduled downtime	Equipment failure causes non-productive downtime. During this time no lot will execute a process. Unscheduled downs are not predictable. Historical analysis is used to compute a downtime distribution to model equipment downtime behavior. The related parameters are the MTTF, MTTR, and the distribution function.
Scheduled downtime	To avoid unscheduled failures, regular equipment maintenance is necessary. Such equipment downtimes are predictable, depending on the particular time horizon. Either the preventive maintenance plan or a down distribution is used to model scheduled downtime.
Current equipment state	The current state defines if the equipment is in a productive, an idle, a down, or setup state. In order to do accurate initialization of the fab, the current equipment state is required.

Table 4: Equipment data requirements

4.1.4 Comparison of Steady State and Transient Simulation Approaches

For planning purposes and operational decision making different simulation approaches are used. This section compares the data requirements for both approaches. For planning, all relevant data contributes to a long term steady state. A detailed initialization is useful but not necessary. The result is a warm-up period, whereby the results deviate from the steady state. This period will be ignored for statistics collection (Robinson 2004). For operational decision making, the transient simulation approach applies. Therefore a well initialized model state is extremely important (Reijers and Aalst 1999). The data needs to capture the current state and the long term behavior of simulation entities. For the near future it is also feasible to use operational planning data if it is available. So the different simulation approaches lead to different model requirements which lead to different data requirements. Table 5 provides an overview of different data requirements.

Model element	Long term simulation for Planning purpose (Steady state simulation)	Short term simulation for operational decision making (Transient simulation)
Future lot release	A historical analysis is used to define future lot release. An average release rate defines the amount of released lots per time unit. This approach has an infinite time horizon.	A lot release plan is used, which contains detailed information about the lot name, the future release time, and additional lot properties. The time horizon of the lot release plan is limited.
Initialization of WIP lots with current operation	It is an option to use WIP lots with the current lot operation. Compared to the starting point of an empty fab, it reduces the warm-up period to reach steady state faster.	WIP lot initialization is essential for short term simulation. It is necessary to obtain information about WIP lots, their current operation, and the current equipment.
Remaining process time	It is not used.	This parameter is used for lots which are processing at the simulation start. The remaining process time is used to determine more precisely when the lot finishes processing.
Lot exceptions	Long term data for rework, hold, sampling, and split/merge is required.	Long term data for rework, hold, sampling, and split/merge is required. In addition to that, also the initial state information is required.
Unscheduled down	Long term statistics are required. A historical analysis generates MTTF and MTOL values for a downtime distribution.	Long term statistics but also the current state is required. It is necessary to capture the current equipment state and its expected downtime.
Scheduled down	Long term statistics are required, similar to unscheduled down.	Long term statistics and the current state are required similar to unscheduled down. In addition it is possible to obtain future preventive maintenance information for a limited time horizon.

Table 5: Data requirements for long term and short term simulation

4.2 Data Concept

This section contains the data concept of online simulation. It provides several options and conceptual decisions to design such a system. The overall objective is to meet the data requirements of online simulation.

The first implementation decision is to define the **degree of automation**. Several options are available to create a simulation model like manual, semi-automated, or fully-automated model creation. To meet the requirements of short term simulation it is essential to follow a full automated model generation approach. The main reason is that the latest data is required. A delay in data transfer should be minimized to meet the "real time" requirements. The data transfer process is a repetitive process which is commencing whenever a simulation starts. The simulation model is a highly detailed model of a large scope. Many long term changes like route changes and equipment set changes have to be applied. So the model requirements, to be up to date, make a full manual update impossible.

One decision is to connect to **one, or to multiple fab data sources**. A decision process is necessary to evaluate which data source(s) meet the data requirements. A good way is a connection to the company's DWH, since most companies already use data warehouse technology. The expectation is that most data is available, integrated, and consistent in one single source. The analysis shows, that a lot of data, which is necessary according to the data requirements, is not available within the DWH. The option, to obtain this missing data manually, is not achievable because it needs to be updated daily. Several local database solutions exist with various required information. This data is distributed to numerous heterogeneous data sources. So, for the data modeling, described in this thesis, it is necessary to integrate multiple sources.

Another aspect is to choose between different options to create an integrated database. In database literature several **methods of data integration** are available like a federative approach or data warehouse approach (Lehner 2003). A federative database approach is characterized by a virtual propagation of queries to each sub systems, while the data warehouse approach creates redundant information from its sources. The decision is to use a data warehouse approach, which replicates data physically, to become more independent from data sources and to guarantee the requirement of high data availability. If a data source becomes unavailable, the copy is still available in the data model. The requirement of referential integrity and accuracy can also be reached due to a flexible target schema definition, including primary and foreign key relations.

Another decision is to define **the number of transformation steps**. In other words it is the definition to use different layers. The question is if one step is sufficient or if multiple steps are necessary to transform fab data into simulation model data. A one-step-approach with a direct interface between data sources and simulation model is one option. The disadvantage is that it is very complex, not flexible, and hard to maintain. If some changes apply, the whole interface for import, transformation, and export needs to be fixed, instead of one small part. The source code contains many dependencies between the components for import, for transformation and for export steps. Such a complex system is very hard to debug. So the decision is to use a multiple steps approach to create the simulation model automatically. The functional elements are separated and easy to maintain. Another reason to use this approach is the high number of required data sources. To integrate these heterogeneous data sources, several data cleaning and data transformation algorithms are necessary. The data model is connected to many sources like plain text files, MS Excel, MS Access, ORACLE databases, and Real Time Dispatcher (RTD) repositories. To increase the data quality according to the completeness, the accuracy, and the integrity requirement, multiple options of alternative data sources become available. Another advantage is that the data model is reusable and the simulation software is independent. Multiple different applications like scheduling and simulation optimization are able to connect to this data base (Kohn 2009).

The last option describes the **data transfer** into the integrated database. According to database literature, different approaches are available. Data refresh, data update or snapshot concepts are most common. The data refresh simply overwrites all data. A data update is more intelligent. It just replaces the changes but not similar entries. A snapshot concept replicates the data and keeps a physical copy of them. In the created data model, an update and a snapshot concept have been used. For few very large tables with only minor changes, an update concept has been applied. For most other the tables with regular changes, the snapshot concept has been implemented. It queries the data sources and creates a new snapshot copy to store this data. The snapshot and the corresponding data represent the current fab state. So the integrity requirement within the snapshot is guaranteed. The reproducibility requirement is met because the simulation model can be created from all previous snapshots. The snapshot concept also guarantees a high availability because if one database is not available and the snapshot creation fails, it is still possible to use the last snapshot to create a model.

The conclusion of the design decisions from above is the characterization of the data model in Table 6.

Design Option	Design Decision
Degree of automation	A module is necessary to extract data from the fab and create the simulation out of it. It requires an interface to the fab data and a module to generate the simulation model files.
Number of fab data sources	A data analyzer is necessary to handle data errors, data inconsistencies, and missing data.
Methods of data integration	It is necessary to have a database between the fab interface and the simulation model interface to store the data physically.
Number of transformation steps	The conclusion is to separate the transformation steps. The first transformation step is part of the import module. A basic data transformation is required to import data from a schema from fab data sources into the integrated data schema in the data model. Further transformation steps occur in the database to handle data errors and missing data. The final transformation creates the simulation model out of the database.
Data transfer	Snapshot concept is used to archive multiple physical copies of the fab data. A new snapshot will be created for the most up to date fab state.

Table 6: Design decisions

4.3 Implementation

Within the online simulation project, two data models have been implemented. The first one was used to show that it is feasible to create the simulation model out of this database. The second one has been implemented by an external company as a productive system. The experience made from the first data model directly affects the second data model. The basis of this thesis is the first data model.

The first part of the implementation section contains the developed data schema. The implemented components of the data model are available in the second part. The third part presents the data integration conflicts and solutions.

4.3.1 Data Schema

The purpose of this data schema is to reflect the fab entities and relations. The schema is simulation tool independent. Figure 10 shows an illustration of the core data schema. This illustration is used to show the basic relations and the basic concept. Due to the high complexity of the wafer fab the illustration of the full fab schema is not appropriate.

According to the data content requirements in section 4.1.3., the data model schema contains all relevant information about lots, routes, and equipment. The concept of using future planning data is available for the lot release table and the PM table. The basic concept of a high detailed initialization is also in evidence. In Figure 10, the current equipment state is available in the equipment table. The remaining process time is available in the WIP lot table.

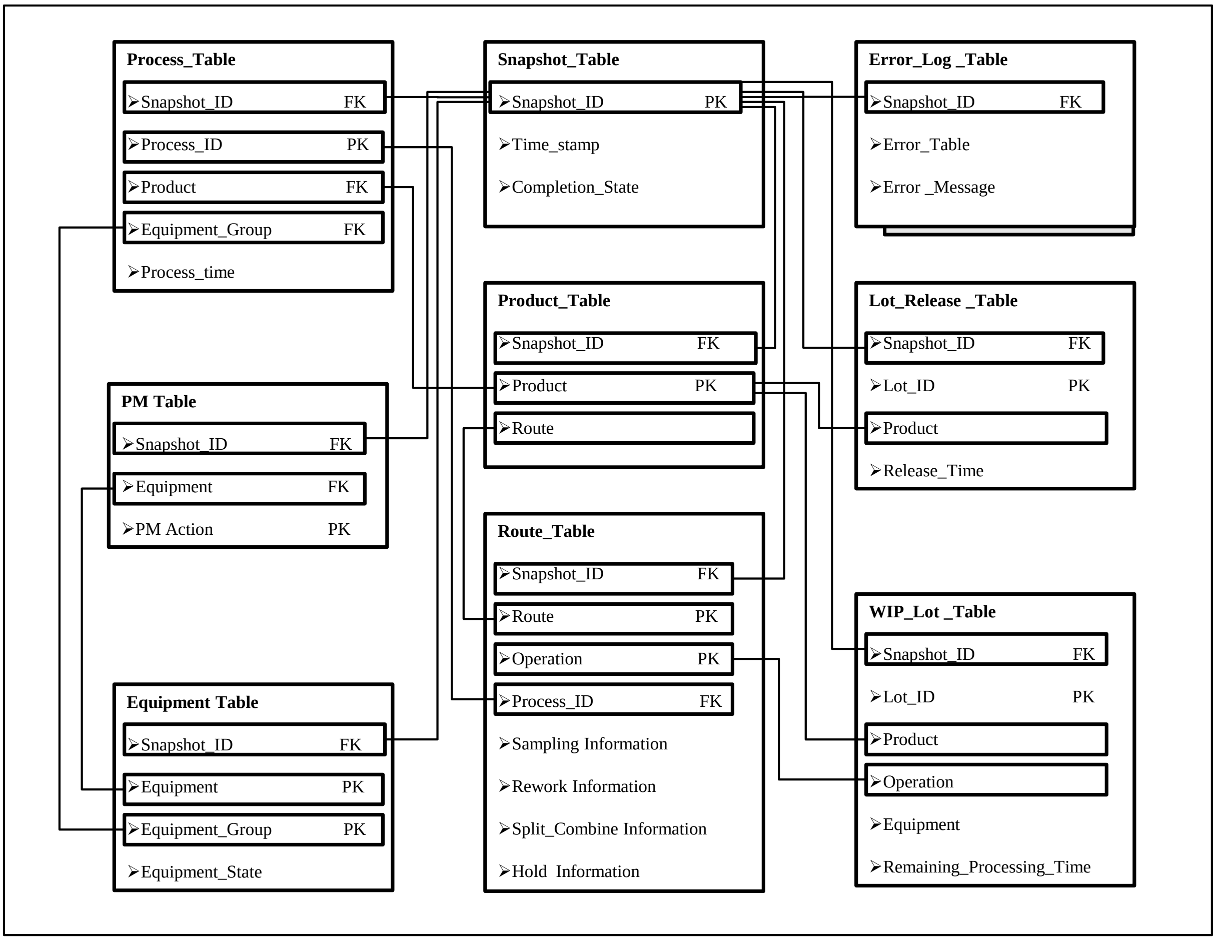


Figure 10: Simplified data schema

To model the relations among entities, the primary/foreign key concept is used. An example for primary key is the product in the product table. An example for a foreign key is the product in the lot release table. No lot can be inserted without a valid product, available in the product table. When inserting a lot without a valid product, the system throws a referential integrity exception. The “delete cascade option” guarantees that all lots will be deleted too, when the product key will be deleted. So, referential integrity can be guaranteed at any time. By achieving referential integrity, the data model avoids simulation exceptions.

A proper snapshot concept is available as well. Each table either has an underlying Snapshot_ID, as an extension of the primary key. It is used to archive multiple historical fab states. The snapshots are organized in a round robin mechanism. The advantage is to avoid storage overflows. On the other hand it is still possible to analyze and rerun historical simulation models and underlying data. For the round robin mechanism, the most current Snapshot_ID will be stored. When creating a new snapshot, the system will iterate the Snapshot_ID and overwrite the oldest snapshot. For the described system the past 90 snapshots are available for debugging. So for the daily snapshot, about 3 months of historical information are available. Every time a new snapshot will be generated, one old snapshot is deleted. The same Snapshot_ID stores all relevant information. The new creation data in the snapshot table indicates the most up to date time stamp of the fab state. Beside the round robin it is also possible to create a new snapshot outside of the round robin mechanism.

4.3.2 System Components

Beside the database itself, many system components are part of the data model. All of these elements are required to support full functionality. Figure 11 depicts the conceptual framework of the following data model system components:

- Data import to extract the data from the fab data sources
- Database to store the master data and the snapshot data in the predefined data schema
- Data Analyzer for complex data transformation steps
    - Historical data analysis
    - Error handling
- Model Generator to create the online simulation model automatically
- Scheduler to trigger the data processing steps

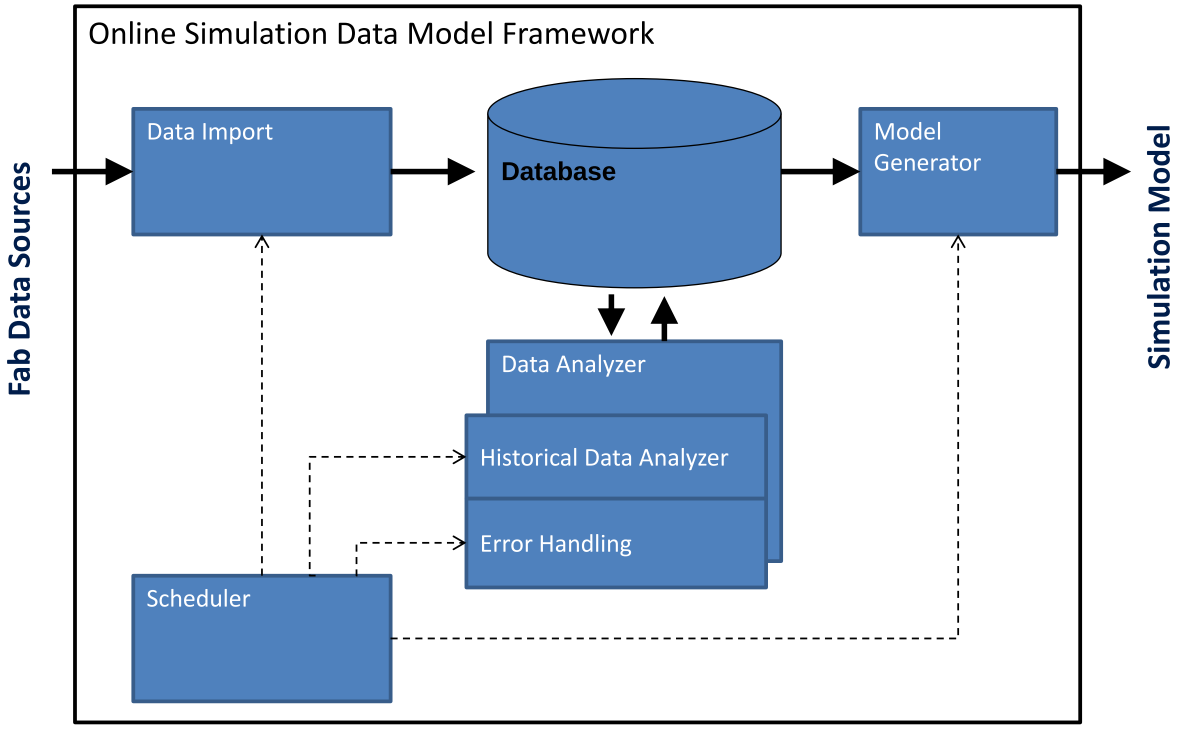


Figure 11: Framework for online simulation data modeling

Data Import

The main task of the data import model is to extract the required data from the fab data sources. Several different data sources are connected. Different applications like text based, MS Excel, MS Access, and common databases are being used. To integrate MS Excel and MS Access, a conversion to the csv format is necessary. If all queries for one snapshot are successful, a completion state is stored in the snapshot table to distinguish, if the snapshot succeeds or fails. The data import module contains numerous transformation steps to map the source schema into the target schema. The major transformation steps are to apply filters, to join information and to apply renaming. Filters are necessary to avoid the transfer of irrelevant data. A renaming is necessary to follow a single naming standard. Joins are necessary to define relations for different data. Examples are:

- Filtering
  - Dummy values
  - Duplicated entries.
  - Values from other facilities
  - Historical data that is not relevant for current production, like old products, routes, or removed equipment
  - Entities which are not in the focus of the simulation, like test lots, or engineering equipment.
- Renaming
  - Synonyms for Equipment: Machine, Equipment, EQ_NAME, EQ_ID, Tool
  - Synonyms for Lot: Wlot_number, Lot_ID, tu_id
  - Synonyms for Product: Wlot_prod, product_numer
- Joining information
  - Merge the equipment with equipment location to generate the transportation matrix.
  - Add related equipment grouping information to equipment.

The additional task for the data import module is to manage the snapshot IDs. It creates a new snapshot ID for every snapshot. The round robin mechanism is implemented in the import module to overwrite old snapshots. This is necessary to limit the required storage space. The data import is implemented in SQL and RTD- Formatter. Therefore the software RTD Formatter Version 7.4.2 and oracle SQL developer 1.2.1 have been used.

Database

The database itself is the central element of the data model. It contains the data values in the predefined data schema (Figure 10). It is an integrated representation of the whole wafer fab. The data schema is simulation software independent.

Data Analyzer

The data analyzer consists of two elements. The historical data analyzer generates statistics from historical values. The error handling module finds and resolves data errors. Both parts are presented in detail.

The historical data analyzer is a major part of the data transformation. It generates statistics from historical lot trace and historical equipment state data. Examples for historical analysis are:

- Average hold rate and average hold duration per operation
- Average split rate per operation, related merge operation, and average number of wafer in a split lot.
- Average sampling rate per operation and attribute dependent sampling rates
- Average rework rate per operation, including the rework route, the first rework operation, and the rework return operation.
- Downtime statistics (MTTF and MTTR) for down distribution.

The second part of the data analyzer is the error handling module. Sargent (2005) pointed out the importance of model validation within different phases of the simulation project, e.g. the data modeling part. The automated error handling module is the implementation of such a concept.

The first component of the error checking module checks if the data queries totally fail. In those cases, an email will be send. Reasons for query failures are unavailable data sources, password/login/path changes, and primary key/foreign key violations. This way of error checking is very important to obtain an immediate response from the system because a single data query failure leads to an interruption of the whole query process and the simulation will not start. The simulation is only able to run with complete information. If one piece is missing, the simulation forecast is not valid. This critical situation would require an immediate response by an administrator.

Another component of the error checking module is the data value checking. This module works as follows. Within a data query, the system generated a test table. If these tests do not match the predefined requirements, the system will generate a warning message and write it in to the error log table. Following examples for specific test cases exist, whereby the parameters are out of their range:

- The rework and hold percentage is not in the range between 0 and 100%.
- Less than 100 PM actions are listed for the next week, which are too few.
- A lot has less than 1 wafer or more than 25 wafers.
- The process time is less than 0 or higher than one day.

Several test cases apply to identify further data inconsistencies:

- The product information is not available for a WIP lot.
- The route information is not available for a product.
- The current operation for a WIP lot cannot be found.
- The required process and the related equipment are not available.
- The process time is not available for a combination of equipment, process and product.

In addition to the error checking part, the data correction algorithms apply as a component of the error checking module. Böhl (2010) presents an approach to improve the data quality for the PM plan. It is not workable to correct each error manually and deliver simulation results in real time. It is necessary to resolve most data issues fast and automated:

- If the current operation of a WIP lot cannot be found, the next available operation number is used.
- If the process time for a combination of equipment, process, and product is not available, than the average process time for the equipment is used.
- If the required process or the required equipment cannot be found, the dummy process or the dummy equipment is used.

For those cases where an automated solution is not applicable, the user still has to correct data values manually:

- If the product number is not available for a lot, a manual product assignment is necessary.
- If a route is not available for a product, a manual route assignment is necessary.

Model Generator

The automated model generator creates the simulation model out of the integrated database. This part has been contributed by an external company. The model generator is application specific and software specific. Further transformation is necessary to change the data model schema into the meta model of the AutoSched AP simulation software (ASAP), Version 9.3. In very few cases, adaptions of the ASAP meta model also become necessary. The ASAP has been modified by using custom extensions. An example is the process time. In ASAP the key for the process time is the route, the operation, and the equipment group. The process time is stored in the route table. With the implemented extension, the key for process time is the equipment, the product, and the process. An extra table stores the values for the process time. This table schema reflects reality best. The process time becomes product dependent. The new approach also avoids redundancy, because the operation number is not a key any more. In the earlier version a process time change requires to change the process time for each operation, even if the values are the same.

Scheduler

The query scheduler is a very important element to trigger the data processing steps. It executes all queries sequentially. The APF Activity Manager Software, version 7.4.2 is used. It provides numerous functions, like time trigger, schedule, wait until, and/or logic, import/export functionality, and execution of external programs via command prompt. Figure 12 depicts a part of the scheduler. This part executes the queries within the data import module at a predefined time period.

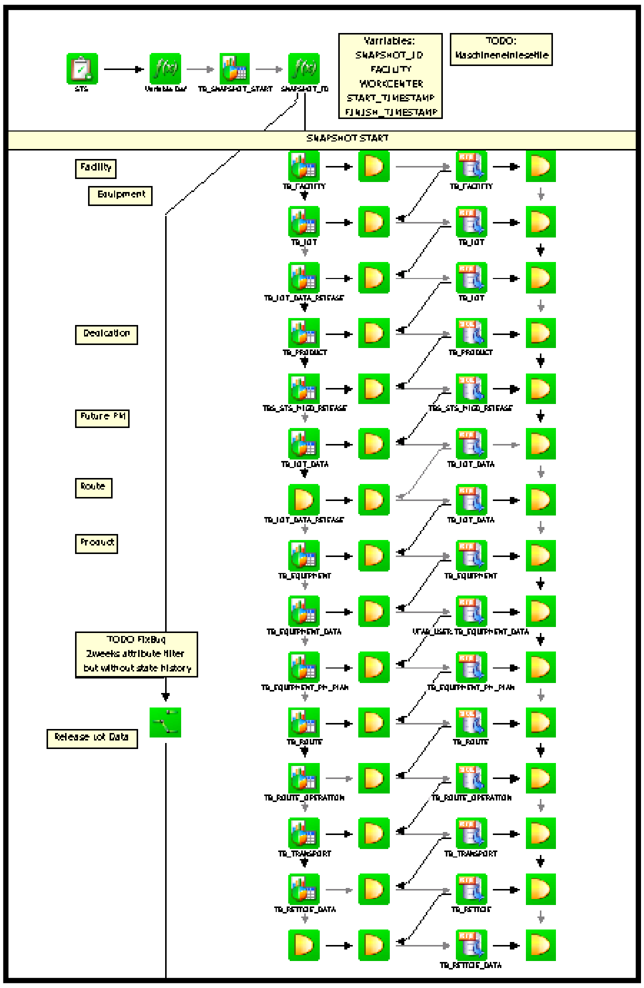


Figure 12: Activity manager to schedule queries

4.3.3 Data Integration

The data integration aspect is important for the implementation of the data model. As described previously, multiple data integration conflicts exist in database literature. So this section focuses on data integration conflicts in the context of online simulation. Several real life problems are present. This is important because the first step to solve a problem is to describe the problem precisely. In addition to that, for each data problem an appropriate solution is available.

Naming conflicts between attributes from different source schemas are quite common. According to database literature, naming conflicts are typical multi source schema problems. Examples are synonyms for attribute names, like: "Machine, Equipment, EQ_NAME, EQ_ID, or Tool". These terms represent the same object. To solve this problem, a target data structure and a project specific naming convention has been defined. Only one term is used to identify those objects. This naming convention is very close to the simulation tool language. It is also close to the common terms in the wafer fab environment. For this particular example the decision was made to use word "EQUIPMENT".
Closely related problems are homonyms. For examples in one data source, a data column is called "Equipment" but it describes different aspects of equipment. It contains data that represents the whole equipment but also the equipment components like chambers, loading stations, lift buffers, and stockers. In most other data sources, the name equipment only refers to the equipment mainframes. The name equipment is misleading. In this particular case the term is not very precise. It represents a different definition. This is a table- attribute naming

conflict. The meaning of this column becomes obvious by looking at the corresponding data. This problem has been solved, by using a clear definition for each attribute name. For the example above, only the entries for the equipment mainframes are used. They are demanded by the process steps in the simulation model.

Other problems are **duplicate entries** in the integrated preventive maintenance (PM) plan in the data model. This is a data value problem caused by having multiple data sources. In each data source this entry is unique, but in the integrated data base it appears twice. One data source contains all PMs for one department. The second data source contains equipment downtimes if an external company is required. So if a department requires an external company, the entry appears twice. In this particular case only unique entries are allowed. The system deletes duplicate entries.

In addition there are several **redundancy** problems in data sources, like the lot release plan. For one group of lots, a range of lot names is available. Another column provides the amount of released lots for that range. The information of the amount of lots appears twice because the name range already represents the lots. An example is "LOT ABC033, "LOT ABC034" and "LOT ABC035", from Table 7. It is a single source, schema and redundancy problem. In this particular case, the system ignores those lots, which have such inconsistency. This type of error occurs due to manual typing errors. In practice it happens rarely.

Another problem is the **precision inconsistency** between multiple data sources. For manufacturing two different sources for lot release plans exist. One lot release plan is structured with lot releases per day, as seen in Table 7. The second lot release plan provides only information about the number of wafer releases per product, per week, as seen in Table 8. The data accuracy is different. This is a schema conflict between multiple data sources.
The option is to decrease the accuracy for the detailed plan, or to increase the accuracy for the abstract plan. The decision is to increase the accuracy to avoid an information loss.
A lot instantiation algorithm enhances the level of detail for the abstract plan virtually. It generates the missing data, for example a virtual lot ID and a virtual start date. It also mimics the typical lot releases behavior during the week. Another precision inconsistency problem exists for planning data sources. The precision of future data is time dependent. Due to an ongoing updating process PM planning data is more accurate if the time difference between the current date and a PM action is smaller.

Release Date	Lot Name Range	Lot Amount
03/05/2009	LOT_ABC033-035	3
04/05/2009	LOT_ZZZ11	1

Table 7: Lot release plan

Release Week	Product	Wafer Amount
34	39847	456
35	23453	233

Table 8: Product release plan

Unit conflicts also occur quite often, especially in statistics. In Table 7 and Table 8, an amount of wafer and amount of lots is available but incompatible. The "lot size" factor converts these values between lot amount and wafer amount. The "lot size" factor represents the number of wafers per lot.

Test or **dummy values** appear in several tables. One example is the route "999999" or the equipment "test". Obviously, it is not used on the shop floor. It is a data value problem within one data source. This requires a cleaning step, before the data import of the route names is executed into the data table. To do so, only those routes are used in the simulation model, which have at least one lot in the current WIP or lot release table.

A common problem is a **different formatting**. The typical example is time format, especially between English and German formats (Table 9). The different formatting is a schema conflict between multiple data sources. As a solution only one target time format has been defined. This format "MM/DD/YYYY hh:mm:ss" is the same as in the simulation model.

Description	**Example**
English without leading zero	9/3/2009 10:19:56
English with leading zero	09/03/2009 10:19:56
German time format	03.09.2009 10:19:56

Table 9: Time format

The most common problem is to deal with **missing data**. According to Skoogh and Johansson (2008) the missing data also has different categories. The categories are that data is available, data can be derived with additional effort, and data is neither available nor collectable. Missing data is also a potential obstacle on the way to referential integrity. An example is, if there is only one productive piece of equipment available without process time existing for this particular equipment. Another problem exists if the complete information is missing. An example is the hold statistics for each operation. An even more critical aspect of data availability is the time dependency of data availability. Future PM's and lot release plans are available for a limited time only. The time horizon for lot release plans is alternating as well. It will be added on a weekly basis. Several techniques have been applied to provide a solution for this wide range of data availability problems. The use of statistical analysis makes the hold, sampling, split and rework statistics available. The replication of previous release plans ensures the lot releases for the full simulation horizon. If many data sources are available for similar information, the most complete data sources have been selected. An example is process time information which exists in three different data sources. Another problem is that a few values are not available at a high level of detail. In this case aggregated data has been used, to replace these missing data values.

A **schema inconsistency** within a single data source exists in the manufacturing system itself. In some areas the dedication does not occur on equipment level but on chamber level. Unfortunately most other tables with equipment information do not contain chamber information. The solution is to use these chambers as pieces of equipment while all other tool models are on equipment level. To obtain all information for these chambers, the information from the equipment has been mapped to these chambers.

Another conflict is the data **value contradiction** between different data sources. For process times or batch sizes, many different data sources exist. The problem is that it is not clear which value is correct and which value is wrong. As an example the process time values for different kinds of equipment differ very much. Furnace and wet bench processes are very consistent due to constant process times. Process times for cluster tools are highly inconsistent due to several process time dependencies like slow-downs during parallel lot processing, reduced process speed during chamber down states, and small lot sizes. The solution is that this data which highly fulfills the completeness criteria has been used. As a first step of improvement the discrepancies between the data sources have been highlighted. The target is to address this issue to the staff who is in charge of the values in the data sources.

Massive data conflicts exist regarding the **correctness** of data values. This is a critical condition, especially when only one data source is available. As an example from the real fab, the data values for the equipment throughput have several correctness issues. The effect of the correctness on the simulation results is significant. The following example of simulation forecast results for the WIP of a single work center demonstrate this effect. Figure 13 and 14 contain results with incorrect throughput data. Figure 14 shows WIP results with corrected throughput data. For the particular example the original throughput of the work center is 13 min per lot. This value comes from the fab data sources, maintained by the production department. In reality the lot trace indicates a value of 10 minutes per lot. To exclude other reasons for the WIP increase, the work center has been analyzed, whereby it turns out that the throughput data source is the reason for the WIP increase. The production department agrees that value from throughput data source is not up to date.

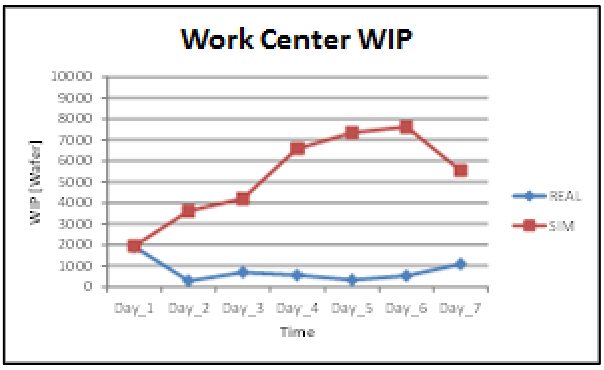


Figure 13: WIP with incorrect throughput

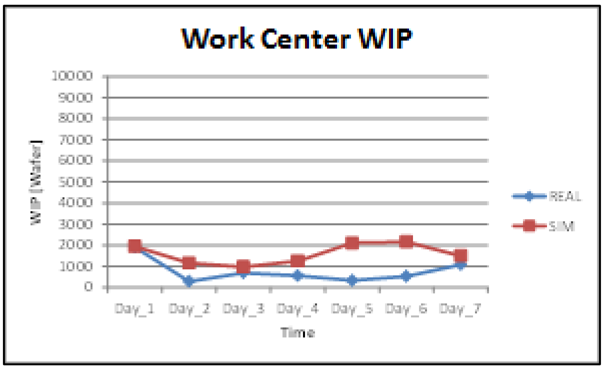


Figure 14: WIP with correct throughput

It can be seen that the effect on work center WIP in simulation is high. In Figure 13 a long queue built up at this work center. The work center becomes a major bottleneck in simulation whereby in reality it is not. With the corrected values in Figure 14, the work center WIP is much closer to reality. This problem is hard to solve because the problem size is large. For the full fab model with a high level of detail more than 100000 throughput time values exist. It is also difficult to compare the data values, for example department interviews, or alternative data sources.

The problem is that it is not clear which value is correct. Data completeness issues do also exist. Another problem is that a small time change of only a few minutes, like the one in the example above, has a huge effect on the simulation results. Altogether it is hardly possible to predict those problems before running the simulation model.

To solve this problem of data correctness, it becomes necessary to create and run the simulation model to figure out, where the problems are. The data validation part becomes a part of the simulation model validation. Based on simulation results it is necessary to figure out, where the data problems are. To retrieve the correct values, a manual lot trace analysis and various department interviews have been done during the simulation model validation phase. An automated, fab wide solution is hard to achieve, due to several dependencies and constraints of the throughput modeling. Frantsuzov (2011) shows, that it is possible to generate data with sufficient quality for some equipment, but not for all equipment in the fab. Therefore this approach does not meet the completeness criteria. For this thesis about 50% of model validation reasons are caused by incorrect input data, which is very much. A lot of effort has been made, for the combined data validation and the simulation model validation. The major reason for incorrect data is that the data is manually maintained. In many cases, the data is not up to date. Furthermore ongoing changes arise in the fab, whereby the new data is not available fast enough.

4.4 Data Model Conclusions and Outlook

All together the data model meets the requirements. It is feasible to create a simulation model out of the data model. To implement online simulation, the data model represents the whole wafer fab. All relevant data for route, product, equipment etc. is available. The data quality improvement is an ongoing issue.

4.4.1 Data Model Facts

The data model contains about 21 tables with about 4 GByte of data. It stores about 100 snapshots. The daily snapshots fill about 16 data tables. It takes about 7 minutes to create a new snapshot. Further improvements reduce this query time. Five master data tables like the equipment dedication tables are also updated. In total, 12 different data schemas are connected to the data base. The data base for simulation imports data from 60 different tables within these schemas. There are still many ways to improve the performance. For this thesis, the focus was to test the data access and the feasibility of an automated model generation for short term simulation.

4.4.2 Data Model Conclusions

From the **data content** perspective, all relevant modeling data is available. If the data is incorrect or not available in the fab databases, a manual maintained solution applies. Such manual data entries are reduced to a minimum. They represent less than one percent of the data.

From the **data quality** perspective much effort has been made to transform the data according to the model requirements. The automated error handling and error correction algorithm increases the data quality very much. Still the data quality issue is a major aspect of the model validation. In many cases it is not achievable to correct data automatically, especially incorrect data. The effect is that the reduced data quality affects the quality of the simulation results. It requires massive manual effort to identify and fix it manually. It is recommended to continuously monitor the data quality on the customers side, to ensure a high data quality (Pipino et al. 2002).

The data model has reached a high degree of **automation**. The system schedules the whole process. It is running without user interaction. The whole online simulation process has been automated. The automation routine includes the following steps:

- Query the fab data
- Transform and store the data
- Create a simulation model
- Run the model
- Generate the results

The **availability** of the data model system reached a level of more than 99%. The main success is that the availability increased a lot. Many difficulties, which cause data model unavailability, have been solved. Examples see below:

- Primary key violation
- Data model schema changes
- Execution time improvements
- Login, password and path changes
- Simulation server and database server maintenance down
- Source database down

Such systematic errors, like primary key violations have been fixed. In the future a parallel development on a productive system will not happen. Therefore schema changes are also not an issue. Regarding login, password, and path changes, it is necessary to keep the data access up to date. Other reasons, like a server down or source database downs, rarely happen. The expectation is that the availability will remain at a level of about 99%. The precondition is that the system will be well maintained.

The data model is capable of **reproducing** a simulation model from a historical fab state. The snapshot concept makes it possible to store a lot of manufacturing data from historical fab states. Therefore the model validation for a historical point in time becomes possible. It is also feasible to generate and compare simulation models from different points in time. If a problem exists, it is possible to determine if the error source is the simulation model, the data transformation step, or the raw data source.

Another criterion is the **maintenance** aspect. The key part, to ensure maintainability, is the custom software documentation. It contains the description of the general concept, the data schema, a detailed description for each table and each attribute, the data sources, and the major transformation steps. Because of the fact that during this thesis the stage has not yet been reached, detailed metrics for this criterion are not yet available.

The data model also meets the **speed criteria.** The simulation results are generated before the results are outdated. The data for the simulation model is available within minutes. The concept of master data and snapshot data makes it possible to remove time consuming queries and transformation steps from the critical path. The generation of the snapshots is fast and it captures the most up to data fab state. The master data rarely changes and the frequency of time intensive updates is very low. Further time improvements are applicable.

4.4.3 Outlook: Fab Driven Simulation

For the use case without direct human interaction, the online simulation meets the speed criteria. It provides an automated forecast. The scheduler triggers a new simulation run and the results are available within minutes. For the use case with direct user interaction, the real time objective has not yet been reached. To run what-if scenarios, the user is not willing to wait several minutes until he gets a well initialized simulation model. To implement online simulation with direct user interaction, it is necessary that the user does not experience extended waiting times. The objective is that the user is able to run the simulation only a few seconds after he triggers the start.

The obstacle is the high initialization time. It includes the time for data extraction, for data transformation, and for model generation (Figure 15 and 16). This duration is not scalable. The time for data extraction and model generation increases, if the model size increases and if the level of detail goes up. Nevertheless, a well initialized and highly detailed model in real time is essential for short term simulation. A solution to increase the initialization accuracy and to decrease the initialization time in the fab driven simulation approach, see Figure 16.

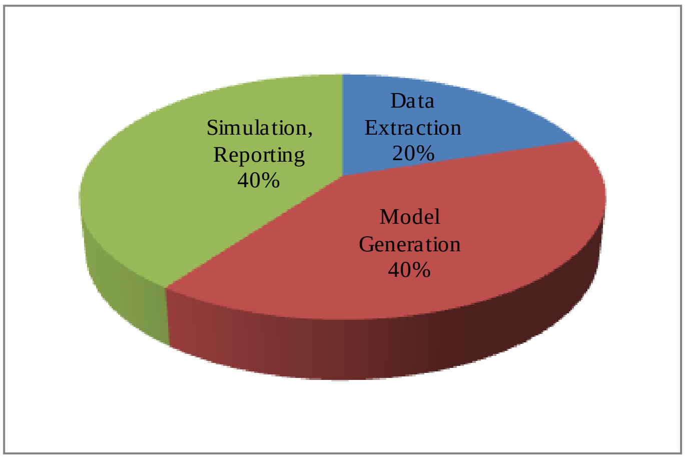


Figure 15: Timing proportions

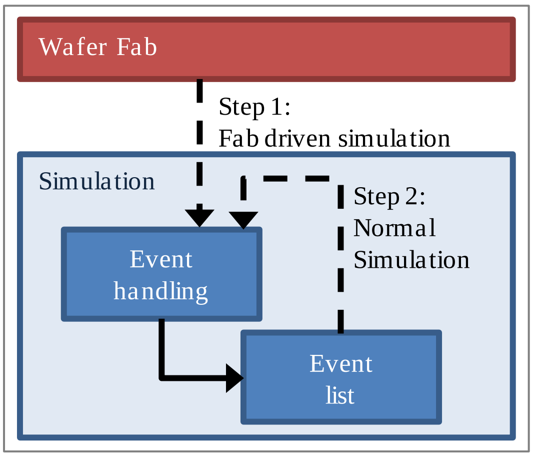


Figure 16: Fab driven simulation

Once initialized, the simulation does not execute the model events from its own event list. Instead, the wafer fab will feed the simulation with all necessary events. It is comparable with a continuous model update within the simulation runtime. By switching from fab driven simulation (Step 1) to normal simulation (Step 2), the event list will provide all events and the wafer fab will not do it anymore. A simulation run for a particular scenario is available instantly. The advantage is that the model itself is up to date, anytime during simulation runtime. The time period for data extraction and model generation, as in Figure 2, is not required any more. Thus, the model becomes scalable. The model size and the modeling details do not highly affect the initialization time, if the fab driven simulation approach is used. The big challenge employing fab driven simulation is the synchronization the simulation model and the real wafer fab. The documentation and prototype implementation of this new idea is available in chapter 8.

5 Simulation Modeling

This chapter describes the elements of the simulation model. It contains the model classification, the requirements, the modeling concept, the executable model description, the validation & verification process, and the modeling facts. The main focus is on the model initialization. The overall modeling results and the achieved accuracy will be discussed in the next chapter.

5.1 Requirements

The modeling requirements are derived from the application requirements. The first part summarizes the general application requirements and defines the required level of detail. The second part contains the basic elements of online simulation. The third part contains the required elements for the forecasting system.

5.1.1 Requirements for Operational Decision Support

The simulation modeling requirements highly depend on the underlying application for online simulation. The generated parameters, their required levels of detail, and their forecast horizon define the model requirements. In the same way the modeling requirements define the data requirements.

The target forecast parameters for short term simulation are work center WIP, lot arrivals, and wafer moves. Most parameters are closely related and require further modeling constraints. The work center arrival parameters need an accurate average lot cycle time modeling. Lot cycle time modeling consist of all lot timing elements like process time, queue time, transportation time, time for rework, setup delays and hold.

For a work center lot arrival forecast, several aspects are required. It is necessary to model the cycle time of a lot, as described above. With a correct average lot cycle time, it is possible to determine when a lot arrives at the work center. Further elements like detailed sampling rates are required to determine, if the lot executes an operation at the work center.

To model WIP, it is necessary to model the elements that have the most influence on the WIP. The WIP basically consists of two elements, the lot arrivals and the lot departures. Lot arrivals have already been discussed. The lot departures basically depend on the throughput of the work center. Several modeling aspects are in the range of interest, like process time, parallel processing, dedication, dispatching rule, and setup. The equipment model like single wafer tool, cluster tool, and batch tool also affect the work center throughput.

For each parameter several time constraints apply. It is required to have a forecast horizon of at least one week. The scale unit for time is days. Due to the short time horizon it is essential to cut the warm-up period and initialize the model very well.

A high level of detail is required to increase the precision and the accuracy of the simulation model. Simulation results become more precise by adding more modeling features. By increasing the level of detail, the deviation between forecast and reality becomes smaller.

5.1.2 Required Model Features

Basic modeling features like route, product and resource modeling are essential for every simulation model (Law and Kelton 1999). For semiconductor manufacturing basic elements like rework, sampling, hold or transportation are relevant too. Many of those elements are described in literature and will not be elaborated any further. All basic and custom modeling elements are connected to simulation output requirements.

Equipment Modeling
For simulation modeling, a proper equipment modeling is essential (Law and Kelton 1999). It consists of the following elements:

- Process time modeling
- Throughput
- Equipment up and down states (Semi 2003 b).
- Setup time
- Additional resources like reticles
- Equipment model types (batching, single wafer processing, cluster tools)

A basic equipment model consists of two elements, the process time and the throughput. The process time does delay the lot. The throughput effects the queue time of the lot. So it directly affects the work center departures, the work center WIP, and the lot queue time. Further equipment model elements like dedication, setup, and equipment model types increase the level of detail. All these elements affect the lot departure pattern of the work center. This is of further importance for a detailed modeling of the lot arrivals for the preceding work center.

Dispatch Rules
Dispatching rule modeling is highly important to model the fab behavior (Sivakumar and Chong 2001). The dispatching rule requirements for this project include following elements:

- Priority of a single lot /product
- Target cycle time consideration
- Time window constraints
- Local dispatching for different work center
- Setup avoidance
- Batching

Without proper dispatching, the executed lot sequence is different from that sequence in reality. Due to the high product mix, the arrival pattern in the preceding work center is also different. This causes a different lot arrival, a different lot cycle time, and a different work center WIP. Without proper dispatch rules, some products are dispatched faster and some products are slower than in reality.

Disturbance Modeling
In a mature high mix wafer fab disturbances occur on a regular basis. Those disturbances interrupt the regular production process and increase the variability of the wafer fab. Without proper disturbance handling, the simulation model is much too optimistic (Randell and Bolmsjö 2001). A typical wafer fab model already handles typical disturbances like equipment downs, rework, setup and hold. In the fab many other disturbances exist.

The reasons for these disturbances are often cost saving aspects or process constraints. Several examples are available:

- Non-productive lots
- Irregular lot release
- Send-ahead products
- Extended dedication rules
- Process time constraints
- Operator behavior

Two examples from above are chosen to illustrate some of the disturbances. The following part describes the send-ahead wafers and the non-productive lots. It is necessary to investigate a disturbance element before it can be integrated into the online simulation model.

Several lots in the wafer fab are non-productive lots. They require equipment resources like any other lots. They are used for equipment monitoring, equipment test, equipment qualification, or development of new products and new processes. Examples are:

- Engineering or prototype lots – basically to create new products.
- Qualification lots- to enable processes on equipment.
- Test lots – to test and monitor equipment.

The main characteristic is that those lots do not strictly follow a standard process route. The lot behavior is not predictable because of arbitrary lot holds, and because of manual interaction. Those lots often need very long measurement times. The lots also enter and exit routes on different operations in the middle of the route. Therefore the desired forecast parameters do not consider these lots directly because they are not predictable. Nevertheless these lots do consume equipment capacity and by doing so they affect the equipment throughput, the work center WIP, and the queue time of productive lots.

The second example is the send-ahead strategy, which is used in the lithography area. The basic principle is to apply a process to one wafer out of a set of many wafers. If this process is successful for the single wafer, it will be applied to all wafers in the set. The benefit of a send-ahead strategy is to reduce the risk of rework and yield loss. Send-ahead wafer and send-ahead lot strategies exist. A send-ahead-lot executes the process step and the related measurement operation to check if the process step is successful. Other lots which require the same process queue will wait until the process information is available. Send-ahead wafers are similar to send-ahead lots, whereby only one wafer in the child lot executes the process step and the measurement step. During that time, the mother lot is waiting. When the process step is successful, the mother lot will execute the same process step as the child lot. The mother lot and the child lots merge afterwards. In the simulation model, the wafer send-ahead modeling feature is incorporated. Department interviews indicate that wafer send-ahead has a higher impact in manufacturing.

Both examples illustrate that many disturbances exist. These disturbances are not predictable. They increase the variability of the wafer fab.

Initialization
A major model requirement is the initialization of the fab model. The simulation model has to be warm started with the current fab state. The current fab state includes the following elements:

- Current lot position, including the operation number
- Lot processing state if the lot is in processing.
- Lot hold and rework
- Lot attributes such as lot identifier, priority, cumulated cycle time, and lot size
- Current equipment state

According to Reijers and Aalst (1999) the short term simulation results highly depend on the initial state. The initialization affects all forecast parameters within the first time horizon of the simulation run. It is not feasible to cut off the warm-up period in a transient fab model. The forecast results proved necessary from the first day onwards.

5.1.3 Simulation Software Requirements

Besides the contents of the simulation model there are several other software requirements that must be met within this short term simulation project.

Similar to the data modeling part, a **high speed** for the simulation run is also necessary. It is important, to get a forecast very fast, before it is obsolete. A fast simulation is critical for later optimization. The simulation execution time basically depends on the number of executed events and the process time per event. The execution time is becomming longer with the increase of modeling details. The simulation software has to be fast to satisfy the demand of a high level of detail and a high speed.

Another customer requirement is to use AutoSched AP (ASAP) **simulation software** (Phillips 1998). The reason is that ASAP is well established and widely used for semiconductor fab simulation. It is also already in use at the customers side and very reliable to generate fab performance analysis.

A further requirement is to do an **error checking** in the simulation model. The objective is to avoid errors that are hidden and affect the simulation results. It is necessary for debugging and error checking. Therefore ASAP generates messages with different debug levels, classified as error, message, and output.

A proper **documentation** is required for simulation modeling to maintain the simulation solution. It consists of the modeling features. Therefore, the underlying concepts, the implementation and the required data are described.

A requirement with regard to the future development is to keep the model extendable for certain **modeling features**, **scenario management** and **optimization**. If the simulation model provides good forecast results, the expectations and requirements will grow. If the number and the versatility of applications increase many different custom requirements come in place. The simulation model needs to deal with these future requirements.

5.1.4 Requirements Overview

The essence of short term simulation is:

- High speed
- High forecast accuracy
- High level of detail

A high speed is required to enable real time decision making. If online simulation consumes too much time, the simulation results will be outdated and the model will not be extendable to optimization. High forecast accuracy is required to ensure that the results do not deviate from reality. Therefore model validation is highly important. A high level of forecast details is useful for small operational units, like single work centers in order of making operational decisions possible. For an overview of online simulation, Table 10 lists the most important requirements of the desired forecasting system.

Characterization criteria	**Desired system**
Environment	Semiconductor manufacturing
Method	Discrete event simulation
Objective	Operational performance improvement
Applications	Numerous operational applications: Proactive dedication management, Preventive maintenance scheduling, Energy saving standby planning, WIP based sampling optimization
Desired execution time	A few minutes
Model creation	Online simulation approach Automated model generation Model creation from fab data sources
Input data sources	Fab data sources including: ● Current fab state ● Historical statistics and trace data ● Future operational planning data
Model scale	Full wafer fab, with the majority of the equipment and lots.
Minimum level of detail, modeling	Lot level Equipment level
Minimum Level of detail, reporting	Product Work center level
Forecast time precision	Simulation time unit: seconds Reporting time unit: days
Forecast time horizon	Short time horizon The time horizon is seven days. A transient simulation model approach is used
Major forecasting parameter	Work center WIP Work center lot arrival Work center moves

Table 10: Overview of online simulation requirements

5.2 Conceptual Model

The first part distinguishes different modeling approaches for offline simulation for planning and online simulation for operational problem solving. The reason is that these approaches are very closely related. Several classification criteria help to distinguish both approaches. The second part is about the definition of a proper level of detail of the online simulation model. An overview is available to define the level of detail for online simulation.

5.2.1 Simulation Model Classification

Operational decisions range between manufacturing execution decisions and long term planning decisions. For manufacturing execution decisions, short term solutions like scheduling are used. Static planning methods and simulation solutions are used for long term planning decisions. Online simulation, for operational problem solving, lies in between.

The difference between scheduling solutions and simulation solutions has been discussed in literature and is already widely known (Klemmt et al. 2008). This thesis distinguishes discrete event simulation approaches for planning and for operational problem solving. Several classification criteria help to distinguish both approaches. In literature multiple classifications exist to describe several simulation modeling approaches (Law and Kelton 1999). For example continuous and discrete events simulation exists, process oriented or event driven simulation exist. This thesis only describes those classification criteria, where the simulation modeling approach is very different in operational and planning level. Table 11 compares those criteria.

Usually simulation models for planning purpose have a very long time horizon, whereby for operational problem solving the time horizon is rather short. The following classification criteria exist to distinguish both approaches:

- Transient and steady state simulation
- Offline and online simulation
- Stochastic and deterministic simulation

Steady state simulation and transient simulation

A steady state simulation is an approach that requires a stable model behavior to gather statistics results. To guarantee stability, it is necessary to avoid overload and to consider a warm-up period. The warm-up period displays transient behavior. It is not used for statistics collection for steady state simulation models. In contrast, a transient simulation model shows unstable model behavior. The reason is that simulation conditions are changing over time. In reality the wafer fab never reaches a steady state, because ongoing changes apply. Examples are product mix changes or equipment changes.

For planning purposes, like capacity planning, a long term simulation is appropriate. The planning horizon, which is the time from where all information is available and decisions take effect, is usually a long time ranging from several months to even years. Operational decisions usually have a short horizon of only a few days. Planning decisions require simulation results for a steady state. Planners are interested in reliable and stable results over a long period of time. Usually operational decisions have only a short simulation time horizon. So the simulation results highly depend on the initial state (Reijers and Aalst. 1999). This thesis describes a transient simulation. Trend changes over time are also in the range of interest and not the steady state.

Offline simulation vs. online simulation

One way to distinguish modeling approaches is the way of the data input and the model creation. Robertson and Perera (2002) provide an overview of different data input strategies.
For offline simulation, the model creation process is a manual process. A lot of time and manpower is necessary to obtain the model data and create the simulation model. For online simulation the data extraction and transform and load process (ETL) are highly automated. In this case there is also an automated model generation approach used.

According to data coupling a simulation system for planning purpose is usually created offline. A mixture between online and offline simulation is applicable. For operational purpose it is useful to create it online. The reason is that operational control is a repetitive task with high time constraints and high data requirements for initialization.

Stochastic vs. deterministic simulation

Another aspect to distinguish simulation modeling approaches is the impact of randomness.
In a wafer fab many uncertainties exist, for examples exceptions like equipment downs, sampling, or rework. Randomness is necessary to consider the effect of exceptions without knowing the exact future behavior. To do so, multiple simulation confidence runs will be carried out. Random number generators create values for missing variable values. Every simulation run generates different results. For reporting, the average behavior of the results is important. The opposite is deterministic modeling. Every modeling component is defined beforehand. No randomness is required and every run will produce exactly the same results.

Another interesting decision element in simulation is the randomness component. It is necessary to figure out if a deterministic or a stochastic modeling approach is applicable for online simulation. Due to the high complexity of a wafer fab the uncertainty for a long term fab simulation is high. A stochastic approach is adequate. The forecast time horizon to solve operational problems is rather short. Many elements are known beforehand, like future lot releases, initial WIP lot positions, and initial equipment states. This is in compliance with the deterministic modeling approach. The key issue is to find out how big the influence of uncertainty is for a short term fab simulation. For this thesis it will be analyzed further what the advantages and disadvantages of a deterministic and stochastic modeling approach are. The simulation modeling chapter addresses this issue.

General Purpose	Planning	Operational decision support
Transient or steady state	Steady state	Transient
Data coupling	Offline	Online
Randomness	Stochastic	Stochastic or deterministic

Table 11: Comparison of simulation approaches for planning and operational decision support

5.2.2 The Level of detail

An important aspect of the conceptual model is to define a proper level of detail because it becomes necessary to increase the usefulness of a model without increasing the complexity.

5.2.2.1 Example Decisions to Define the Level of detail

The simulation model of the wafer fab is very large and it contains several modeling features. Therefore it is not useful to elaborate on the decision process for each feature. The following part describes decisions to define the level of detail for transportation, process time, and sampling.

The first example defines the level of detail for **transportation.** It is possible to create a fab model without transportation. It is possible to have a low level of detail with a single delay time or to have a simple work center transportation matrix. It is also possible to create a high level of detail transportation matrix on equipment level with a distribution for each cell in the matrix. From a modeling perspective a high level of detail is not very practical because:

- The variability of transportation data is extensive (derived from department interviews).
- The data is not available on the desired level of detail. A historical data analysis is necessary to generate the data. The events for each transport job are the data source.
- Effort for implementation and model execution are enormous. A random value needs to be generated for each transportation event.
- The benefit is small because transportation is a delay time without major queuing effects with regards to the work center arrival and WIP.

Therefore the transportation feature has been modeled on a very abstract level. A transport matrix with the level of detail "building" has been created, like seen in Table 12. The values are not realistic due to non-disclosure agreement.

From\To	Building_01	Building_02	Building_03
Building_01	5 min	10 min	70 min
Building_02	10 min	5 min	70 min
Building_03	70 min	70 min	5 min

Table 12: Illustration of transportation matrix

Another example is the **process time and throughput modeling**. The process time (and batch interval) highly effects the delay of the lot, the throughput of the work center, the work center WIP, and the arrival rate for the downstream work center. This is the major reason to model process time with a high level of detail. In the simulation model, the process time depends on equipment, process, and product (Table 13). For single wafer tools the time per wafer is available to model small lot sizes. For this level of detail, all relevant data is available in the fab. The implementation strictly follows the defined level of detail.

Equipment	Process	Part	Process time [min/lot]	Process time [min/wafer]
Eq_01	*	*	21.5	
Eq_01	Process_01	Part_01	21.5	
Eq_02	Process_01	Part_01	39.1	
Eq_03	Process_01	Part_02		5.3
Eq_03	Process_02	Part_02		5.3

Table 13: Illustration of process time dependencies

Sampling Modeling

In semiconductor manufacturing, the sampling mechanism is used to trigger optional process steps, especially for measurement. Wafer sampling and lot sampling exist. For Wafer sampling not all wafers in a lot execute the measurement process. The wafer sampling affects the process time of the lot. From a modeling perspective, the wafer sampling is being captured by the process time. It already considers this behavior. The lot sampling is more interesting from simulation modeling perspective. Lot sampling means that not all lots need to execute a particular operation. Often only a small amount of lots execute the measurement step. The sampling feature is highly important to model lot arrivals. Without consideration of the sampling feature, the lot arrival forecast for a work center is not usable. The lot cycle time and the work center WIP for the current and the preceding work center are also affected. Therefore it is necessary to apply a sampling concept. The common way is to apply an operational sampling rate. Many simulation tools like ASAP offer this option. So the decision is to use operational sampling rates. During department interviews it turned out that additional dependencies, such as lot attributes, exist. They affect the sampling behavior and are also available before simulation start. The decision is to extend the common sampling concept for a pilot study. It will be implemented for those work centers where lot attributes affect the sampling rate.

5.2.2.2 Decision Criteria

The decision about the level of detail depends on the several criteria. For this thesis, each modeling feature has been checked against the following criteria to define the level of detail.

Increase of simulation accuracy	Those features with a high benefit for simulation will be implemented with a high level of detail. Model features where the estimated benefit is low will not be implemented or will be implemented with a low level of detail.
Variability of the model feature	When the underlying variability or uncertainty for a feature exist, it is not useful to increase the level of detail, as long as the uncertainty cannot be solved.
Data quality	It is clear that good data quality increases the benefit of a modeling feature.
Effort for implementation and runtime	Implementation and runtime effort stands in contradiction to the potential accuracy gain. The increase of the level of detail of modeling features increases the runtime of the simulation model as well.

Table 14: Criteria to define the level of detail

Additional criteria also apply. The level of detail will be increased if a particular model feature is also closely related to an operation application. An example is the dedication feature which is usable for the dedication optimization in the fab because the chance is high, that a particular model feature becomes a major part of an optimization system at a later stage. The model has to be sensitive to this kind of influences.

5.2.2.3 Overview of the Level of detail

The decision criteria from above have been used to define the right level of detail for all simulation components. For the three examples transportation, process time, and sampling rate, the decision process has been described. The bullet list below shows options to define the level of detail for simulation modeling features. According to this list, the executable model has been implemented. The final implementation of the levels of detail is written in bold:

- WIP Lot
  - Initialize the fab model without WIP
  - Initialize the fab with current lot position
  - **Initialize the fab with lot position and remaining process time, including disturbances like rework, hold**
- Lot release
  - No lot release
  - Lot release rate for each product, derived from historical analysis
  - **Release of lot instances**
- Route
  - Abstract route that represents real routes
  - **Real routes**
- Transportation
  - No transportation
  - One fab value for all transport times
  - **Transportation matrix on building level**
  - Transportation matrix on floor level
  - Transportation matrix on work center level
  - Transportation matrix on equipment level
  - Transport time distribution for the transport matrix
  - Inclusion of the transport system as resources with delay and capacity effect
- Sampling
  - No Sampling
  - Historical lot movements determine a sampling rate per operation
  - **Operational sampling rate with additional dependencies**
  - Detailed sampling as in reality - a math formula with many dependencies
- Equipment model
  - Work center level with a predefined number of equipment
  - **Equipment level**
  - Chamber level
- Process times and throughput
  - Delay step with infinite capacity
  - Fixed process time defined by equipment only
  - Fixed process time defined by equipment and process
  - **Complex Process time dependencies**
    - Equipment
    - Process
    - Product
    - Number of wafers per lot

- Dispatching
    - No dispatching, whereby FIFO is used as a default dispatching rule.
    - Standard dispatching rules which are already available in the simulation tool
    - Combined dispatching rules, whereby a combination of the standard rules exists
    - **Customized dispatching rules**
    - Reuse the source code from the fab dispatching rule itself.
- Dedication
    - No dedication
    - **Dedication depends on equipment, process and product**
    - Complex rules with multiple keys
- Downtimes
    - Unscheduled down
        - No unscheduled down
        - **Downtime distribution MTTF, MTTR downtime distribution**
    - Scheduled down (PM, preventive maintenance)
        - No scheduled down
        - **Downtime distribution MTTF, MTTR downtime distribution**
        - Future PM plan with single PM instances
- Additional disturbances
    - No consideration of additional disturbances
    - Average fab performance reduction to consider additional disturbances.
    - Modeling of individual disturbances in combination with a fab performance reduction
        - **Send-ahead modeling**
        - Modeling of non-productive lots
        - **Hold**
        - **Rework**
    - Modeling of all disturbances in the fab
- Initial equipment state
    - No initial equipment state
    - **State per equipment**
    - **State per chamber**
    - State per work center (number of available tools)
- Randomness
    - **Deterministic model**
    - Stochastic model with confidence runs

During the project execution phase, the level of detail does also change. One reason for that are changes of quality in the underlying data. Adding modeling features increases the level of detail and effect the accuracy of simulation results.

5.2.3 Additional Conceptual Decisions

Many conceptual decisions are required before the implementation of the executable simulation model starts. In the previous section, the appropriate level of detail has been discussed. Apart from this major decision many other decisions are relevant as well:

5.2.3.1 Target Definition

One issue for the online simulation model is to define when the target regarding the modeling accuracy has been reached. The online simulation project has a scientific background, whereby the exact results are unknown at project start. The major target is to achieve a good forecast quality for work center WIP and work center lot arrival.

It is furthermore required that the predicted average value and the average trend changes will match reality. The prediction of future trend changes is even more important than the prediction of the average value for work center WIP and work center lot arrivals. The problem is that there is no exact definition for the phrase “good forecast quality”. Nobody can answer the question when the prediction is good enough. Evaluation of the forecast quality will be on hand if the results are useful. Those results are useful when the staff in the production area is using the prediction in their planning and decision process. Therefore the online simulation system needs to be implemented first.

5.2.3.2 Stochastic vs. deterministic

Another important decision for a simulation approach depends on the degree of uncertainty. Depending on the degree of uncertainty, it is useful to decide between a stochastic and a deterministic simulation approach. If all relevant model parameters are known in advance, only a single deterministic simulation run is required to generate the results. If uncertainty or variability exists, for example due to unexpected equipment downs, multiple simulation runs are required (Law and Kelton 1999).

For modeling of a real wafer fab, a stochastic approach is widely used due to a high degree of uncertainty in the fab. The question is, whether the same conditions apply for short term simulation because much more information is available for a short future time horizon. Uncertainty is significantly reduced. The question is how far in the future the simulation model can look without using stochastic components. A possible way is to run the simulation with the current equipment states and do a replanning if major equipment state changes apply. The question is, whether this replanning approach is feasible for online simulation.

In general, most variability in a wafer fab result from unscheduled equipment downs (Hopp, and M. Spearman 2000). To demonstrate this effect, a real fab model has been used with only one deterministic run. One scenario has been created, where each piece of equipment is initialized with its current equipment state. For every piece of equipment in an unscheduled down state, the estimated down time is also available. After this downtime, the equipment will be up and running in the simulation model. In contradiction to the unscheduled downs, the scheduled down modeling is not affected. Only future unscheduled down modeling is not in use. If no further unscheduled downs occur, the expected result is that the model is much too optimistic. The question is, whether the model results deviate too much from reality.

As it can be seen in Figure 17, the Fab WIP is available for the next ten days in reality and simulation. Due to a non-disclosure agreement, the current fab WIP is available in percent whereby the 100% value denotes the real fab WIP on the first day. The real fab WIP with the blue curve is undergoing some fluctuations. The deterministically simulated WIP with the red curve is too optimistic, as expected. The WIP is becoming lower. During the first four days, the WIP is still quite close to reality. After four days, the difference of about 1% is already obvious. For the next ten days the WIP difference is more than 3%. For a steady state simulation, a difference of 3% is acceptable. For short term simulation, which is initialized with the current fab WIP a resulting WIP difference of 3% is large.

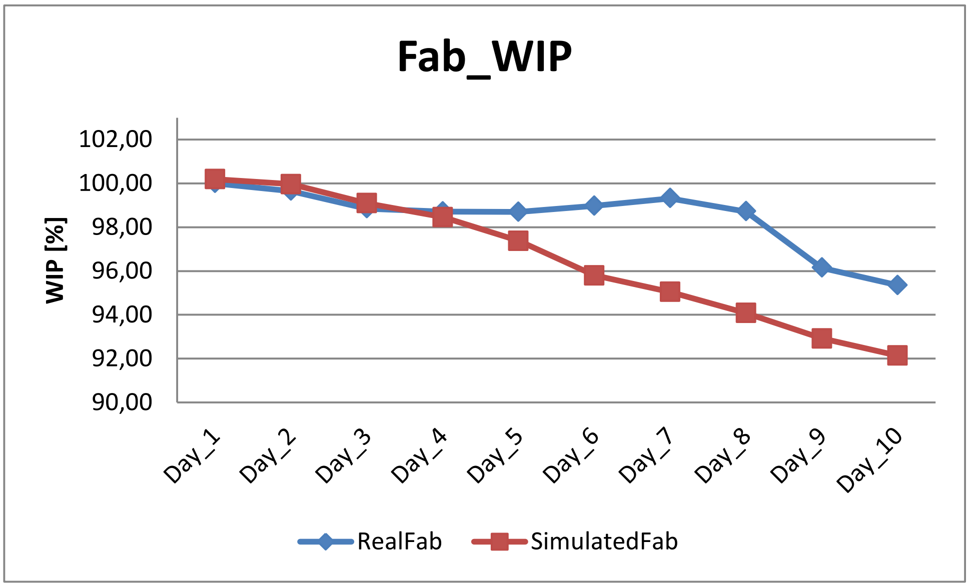


Figure 17: Fab WIP deterministic scenario without unscheduled downs

The conclusion of this short analysis is that the replanning modeling approach is not acceptable because the deviation after four days is already too big. The customer does also not accept such a large deviation on fab level. The requirement of short term simulation is to deliver reliable results for one week's forecast horizon, and not only for the duration of less than four days. The replanning approach, with this configuration, is not applicable here.

The question is still unanswered which approach, a deterministic or a stochastic approach, is most suitable. The quick analysis above shows that stochastic effects are massive and cannot be ignored. The forecast horizon is too long to have prior knowledge about future fab behavior. So the model contains stochastic elements, for example for equipment downs, but also rework, hold, and sampling. The question is how significant the accuracy gain is by using multiple confidence runs.

In general there is a tradeoff between the increase of computation time and the accuracy gain. Is it necessary to execute multiple confidence runs or is just one simulation run enough to make a good forecast? For online simulation it is necessary to consider real time constraint as well.

5.3 Executable Model

The implementation section describes the way, how the custom model features have been implemented. It also contains background information about the reasons, why it is necessary to implement these features. In several examples, the analysis and the modeling results are available to show the benefit of each modeling element.

5.3.1 Equipment Modeling

In semiconductor manufacturing many different equipment types are in use. They are different in the executed processes, the hardware configuration, and the control mechanism:

- Measurement
  - Single wafer process with one lot per process
  - No parallel processes
  - No pipelining
- Furnace
  - Batch process with up to 8 lots in a process
  - No parallel processes
  - No pipelining
- Wet bench
  - Batch process with 2 lots in a process
  - Pipelining of many batch processes running in step sequence through different bath steps
  - No parallel processes
- Lithography
  - Single wafer process with one lot per process
  - Pipelining of many processes running in step sequence of coating, exposure, and development
  - No parallel processes within the critical section because the exposure step exists only once
- Etch
  - Single wafer process with one lot per process
  - Pipelining of many processes when running in step sequence
  - Parallel processing when two or more chambers are available of each step. The detailed properties depend on the custom equipment configuration

In general it is possible to apply an average throughput without consideration of the internal tool behavior. In this case the throughput is defined by the number of lots per equipment. The disadvantage is that the model does not consider the typical equipment properties with sufficient accuracy. Elements that are captured with this approach are:

- Cycle time delay to create a full batch
- Batch criteria, whereby only lots with the same process specification are able to be batched together
- Batch effect whereby many lots with the same batch criteria are released at the same time
- Different sequence of lot departures when lots with small and large process time overtake a lot with long process time in parallel equipment
- Large differences of process time and throughput (time between lot release) when there is a considerable pipelining effect due to a high number of sequential steps
- Increased throughput and process times for small lot sizes

In the simulation model different equipment models are used to consider the equipment behavior with sufficient levels of detail (Scholl et al. 2010):

• Lot process,	No parallelism,	No pipeline,	No batch	(Measurement)
• Lot process,	No parallelism,	No pipeline,	Batch	(Furnace)
• Lot process,	No parallelism,	Pipeline,	No batch	(Wet bench)
• Wafer process,	Parallelism,	Pipeline,	No batch	(Lithography)
• Wafer process,	No Parallelism,	Pipeline,	No batch	(Etch)

It is not necessary to go into further detail for each equipment property. As a single example of such a property, the description of the lot size effect is available for a lithography tool. This effect occurs with single wafer tools like lithography. In the data from Figure 18, the process time depends on the number of wafers in the equipment. The process time to finish all wafers is longer if the number of wafers is higher the in the equipment. There is a linear increase. If the equipment is empty, it still takes time to fill the pipeline, which is called "first wafer effect". This is the reason why the linear curve will not cross the Y-Axis at the process time equal to zero. For comparison, the same relationship is depicted in Figure 19 for a batch furnace. The process time is a constant value. It does not increase, even if the number of lots and the number of wafers in the batch process increases. For both charts, data from the real fab is used. They are based on a historical lot trace.

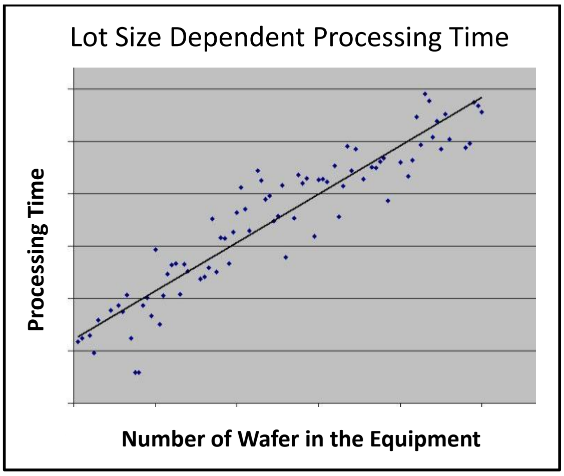


Figure 18: Lot size dependent process time

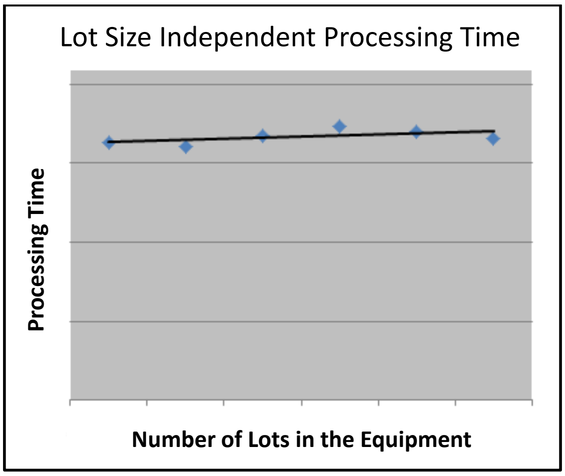


Figure 19: Lot size independent process time

The lot size dependent process time element is widely used, also in the area of online simulation (Radloff et al. 2009, Bagchi et al. 2008). In this area Morrison (2011) pointed out the weakness of such an affine model, whereby the approach of Ax+B is still too simplistic. He pointed out that the deviation from reality becomes large, especially when processing many lots in a row, which have a small number of wafers. The Ax+B approach does not mimic the behavior of the lithography tools very well, if the lot size of all lots in the fab is very small. The cycle time loss and throughput loss is bigger than estimated. The reasons are more setups and also increased lot and wafer handling times for loading and unloading. But in practice most of the lots have full or almost full lot size. Only few lots contain less than 5 wafers. It is more an exception than the rule, if many lots with small lot size are processed in a row. Therefore the Ax+B approach is sufficient for the purpose of fab simulation. It is not useful to extend the level of detail. The effort to include a detailed model of the internal equipment behavior is too high. The time for data collection, model creation, and computing becomes longer.

5.3.2 Scheduled Down Modeling

The most promising approach for this thesis is to integrate the PM (preventive maintenance) plan for future equipment downtime modeling. The PM plan contains the point in time and the duration of a future PM. Examples of the input format for such equipment PMs are depicted in Table 15.

Equipment_ID	Start_Time	Duration[hours]
Equipment_01	04/06/2011 03:00:00	3.4
Equipment_31	04/06/2011 01:45:00	4.5
Equipment_12	04/10/2011 06:00:00	35.0
Equipment_42	04/09/2011 04:20:00	6.9

Table 15: PM input format.

Typically a downtime distribution from historical analysis is used to reduce the average throughput of the equipment. The exact point in time is not known. The main advantage of using PM plans is that the exact time of the PM is known in advance.

The success of this modeling approach highly depends on data quality. In this project a lot of effort has been made to integrate and enhance the information from distributed PM data sources. Figure 20 shows the average equipment down percentage for all production departments. It can be seen that the scheduled down percentage in the PM plan does not represent reality. It is much lower compared to reality. In simulation the result is a higher throughput for each work center.

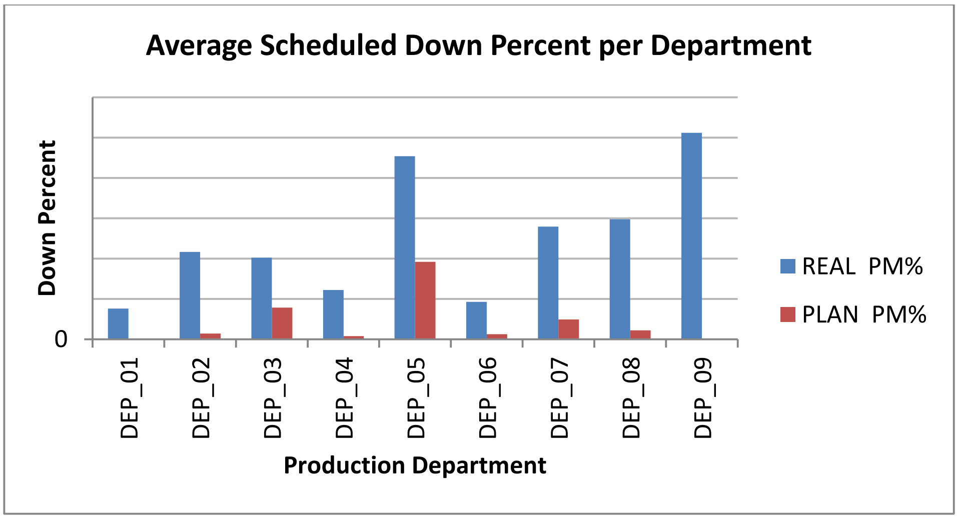


Figure 20: Scheduled down percentage per production department

To figure out the reason for the gap, it is furthermore necessary to keep analyzing the PM plan and the executed PMs in reality. The analysis shows that multiple reasons exist why the PM plan does not represent reality:

- The PM plans contain time based PM but not PMs triggered by wafer count
- Equipment exists where the PM information is completely missing
- The duration of a PM does not contain the time for the test after the PM

The main reason for this behavior needs to be explored. Therefore following numbers affect the average PM percentage:

- Duration of a PM
- Number of PM events
- Number of equipment with PM information

Figure 21 compares the PM duration of plan and reality. Only those events are used for one week, which represents the forecast horizon. As it can be seen in Figure 21, the average duration of the PM plan is similar to reality. In Figure 22 it becomes obvious that the number of events in the PM plan is much too low, compared to reality. Less than 10% of the PM events exist in a PM plan. The number of equipment with at least one PM event is also lower than in reality. In Figure 23 about 20% of the equipment has a PM entry in the PM plan.

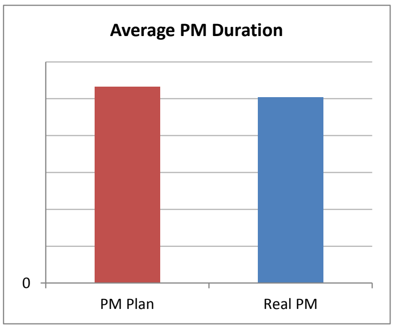


Figure 21: Avg. PM duration

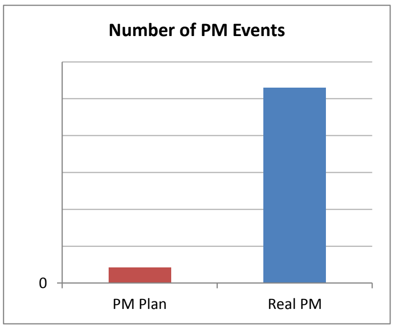


Figure 22: Nr. PM events

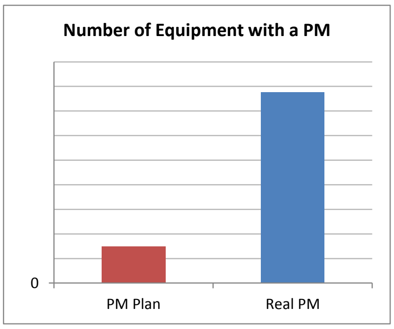


Figure 23: Nr. of PM equipment

The conclusion is that the current data quality of the PM plan is not sufficient. For many pieces of equipment a PM plan is missing altogether. Many events do not exist for those pieces of equipment with a PM plan, especially counter based PMs. In some cases the planning horizon is much smaller than one week. In those cases, the correct PM event is not available before simulation start. To overcome those limitations, the scheduled down modeling has been changed. A distribution is used to model scheduled downs (PMs). The historical analysis generates MTTF and MTTR as input data for the scheduled down distribution. This approach is similar to the unscheduled down modeling.

To see the effect of different PM plans, Figure 24 depicts the wafer moves of the whole fab. It contains three scenarios for different PM approaches. The first scenario shows the fab wafer moves by using no PM information at all. The expectation is that the model is too optimistic, and the throughput of the whole fab and also the wafer moves are too high compared to reality. Second scenario contains the PM plan, whereby only very few PM events will be captured. The third scenario is the used solution whereby historical analysis is used to model scheduled downs with MTTF and MTTR.

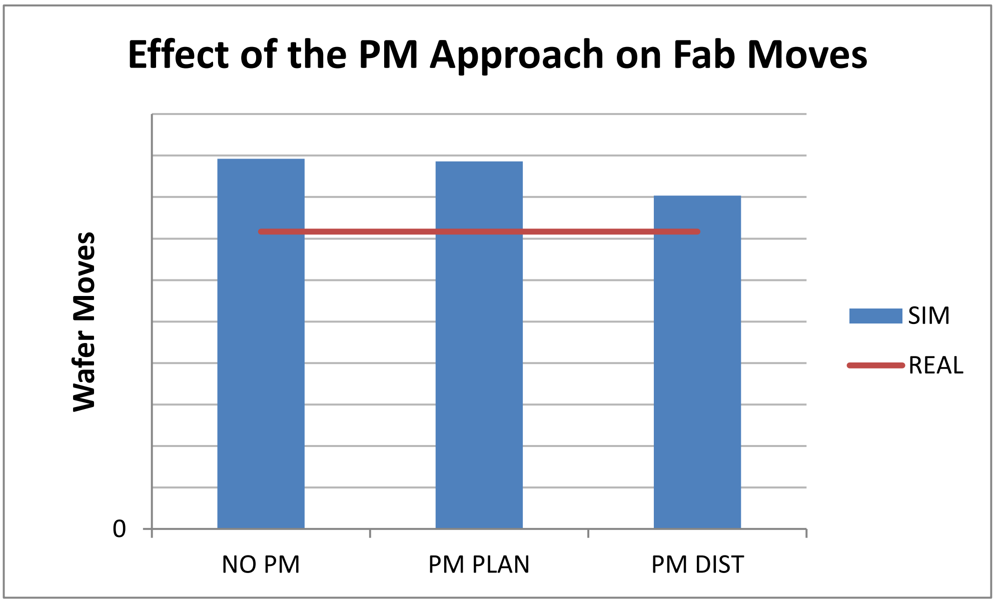


Figure 24: Effect of Lot PM approaches on the wafer moves of the whole Fab

It can be seen, that the modes are too optimistic and wafer moves are too frequent, if no PM is used. The main conclusion is that the usage of the PM plan does not improve the situation significantly. Only the PM distribution affects the wafer moves. They are much closer to reality.

A gap still exists for the third scenario with PM distribution. The reason is not the PM plan itself. The reason is that the whole model is slightly too optimistic. This is because of the disturbances that exist in the real fab, which are not part of the model (Randell and Bolmsjö 2001). The most important message here is that the PM plan causes almost the similar effect as if using no PM at all. Only the PM distribution affects the equipment throughput and improves the model quality considerably.

5.3.3 Unscheduled Down Modeling

For equipment modeling another aspect worth looking at is the unscheduled down modeling. The equipment downs are generated by historical analysis. Due to the fact that the level of detail for simulation modeling is equipment level, the down statistics are captured on equipment level too. The problems are chamber downs, which affect the capacity of the whole equipment. The impact depends on the internal equipment configuration and the internal wafer flow. A chamber down affects the equipment cluster in different ways. Examples are chamber downs for a typical measurement, lithography, wet bench, or etch equipment. For single chamber measurement equipment, the whole equipment is in down state, if the single chamber is down. For lithography multiple coating and development chambers exist, but only one exposure chamber. So if the exposure chamber goes down, the whole equipment is in down state. For wet bench multiple baths are available. A lot will enter a different bath in a predefined sequence. If one bath is down, the lot will be processed as long as a similar bath is available. When no parallel bath is available, the process is not able to run on this equipment. Other processes still run, as long as they do not require this particular bath type. A cluster tool for etch has an equivalent behavior. Different wafer flows are available to process a lot on the cluster tool. Depending on the equipment configuration a single step can be executed on a single or on parallel chambers. The effect of a single chamber down displays itself as follows:

- The whole process on the equipment runs with reduced capacity. The process time increases and throughput will be reduced.
- The whole process is able to be executed on the equipment while another process is not able to run. From lot perspective, some lots are able to run on the equipment, other lots cannot be processed. It is depending on the chamber requirements of the particular lot. It is possible that queue lots exist, which cannot be processed due to chamber downs, even if the state of the equipment cluster is available.
- Due to chamber downs, no process is available for the whole equipment. In this case the equipment state is also down.

One solution is to increase level of detail of the model and add the chamber level. This solution requires to add chambers as resources and to obtain statistics on chamber level. In addition it is necessary to capture the internal wafer flow of the equipment. This whole solution is possible but it is not feasible for a full wafer fab model. The model becomes too detailed. The access to and the overall quality of basic data on such a high level is not very accurate. It is obvious that the model size and the runtime will increase too much.

Another solution is to consider chamber downtimes as equipment downtimes with reduced downtime length. If one chamber is down, it is counted as equipment down. The length of such a downtime is the chamber downtime divided by the number of chambers for this equipment.

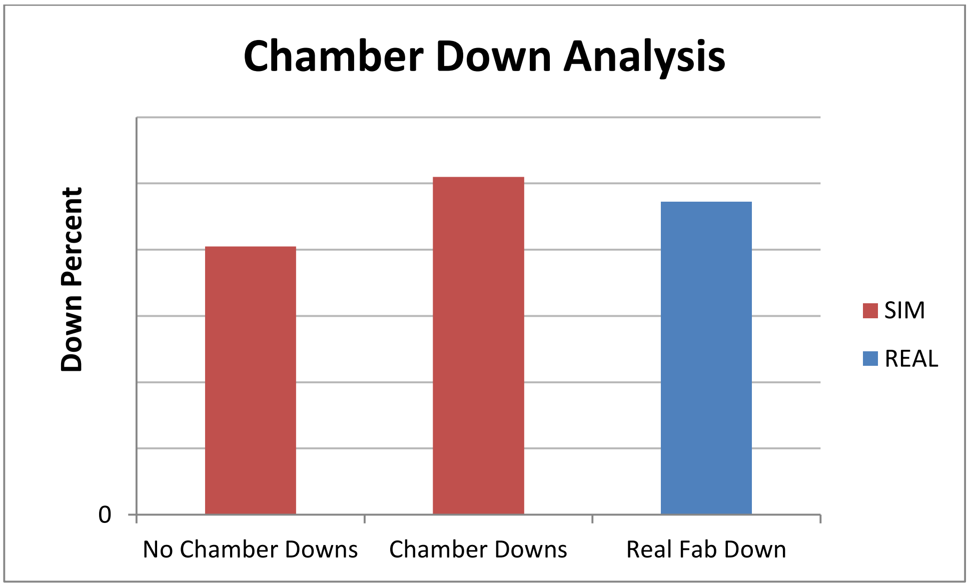


Figure 25: Simulation with and without consideration of chamber downs.

Figure 25 depicts the average chamber downtime of the whole wafer fab. As expected, the average percentage of fab downs increases. The average fab down, with consideration of chamber downs, is higher than without chamber downs. Looking at the numbers behind, it is closer to reality. The gap is reduced to half of its value but it is still too high compared to reality.

As you can see in Figure 26 the deviation from reality becomes bigger, if the simulation model considers chamber downs with this solution. On average the forecast accuracy decreases by 0.5 lots. This approach does not capture the chamber downs on equipment level with sufficient accuracy. The reason is that the internal wafer flow is not known.

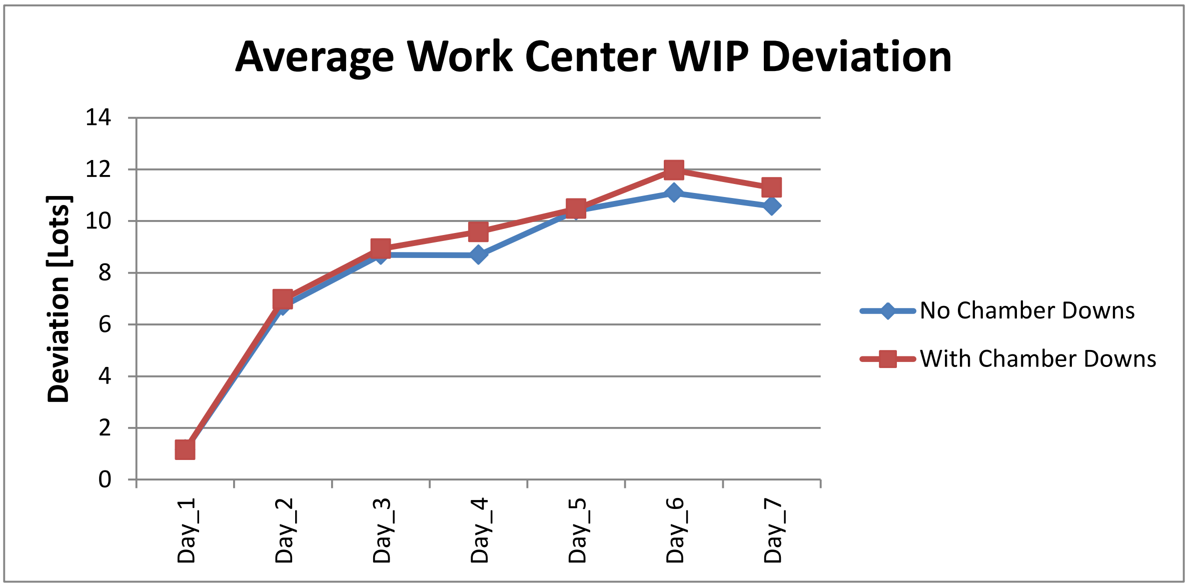


Figure 26: Effect of chamber down approximation algorithm on simulation results

The chamber downs have a totally different effect on equipment throughput. In Figure 27, the first example shows the equipment with 3 parallel processing chambers. This equipment is able to run on one chamber with 1/3 of the throughput. The second example shows a chamber down with parallel wafer flows. It has almost no effect on equipment throughput. The third example shows a chamber down where the whole equipment is affected. For most of its downtime, the whole equipment is not able to run any wafer, even if only one chamber is down.

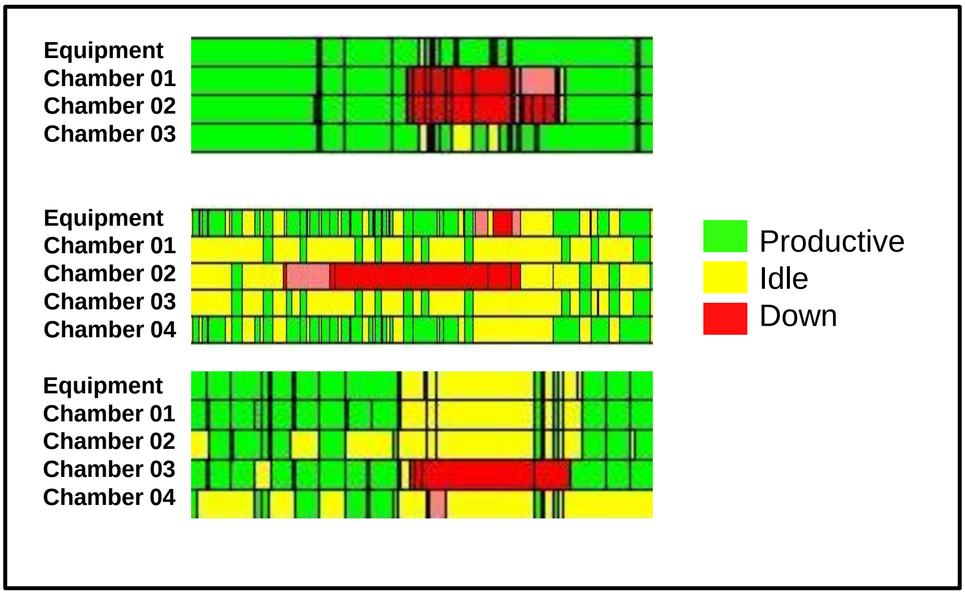


Figure 27: Chamber downs and equipment downs

The solution is to extend the described approach to capture the effect of chamber down for the equipment. The chamber down matrix in Table 16 provides statistics to define this effect. Depending on the importance of the particular chamber down, the effect on the equipment level ranges from 0% up to 100%. It means that the equipment is not affected at all or it is down for up to 100% of the chamber downtime.

Equipment	25%	25%	50%	50%	100%
Chamber_01		X		X	X
Chamber_02	X			X	X
Chamber_03			X		X

Table 16: Chamber up matrix

This information is available in the fab. It is used for equipment downtime reporting. Figure 28 compares the fab down percentage of the real fab with the extended chamber down approach. It can be seen that the down percentage in simulation reflects the real down behavior very well. The difference of the down percentage is less than 1%.

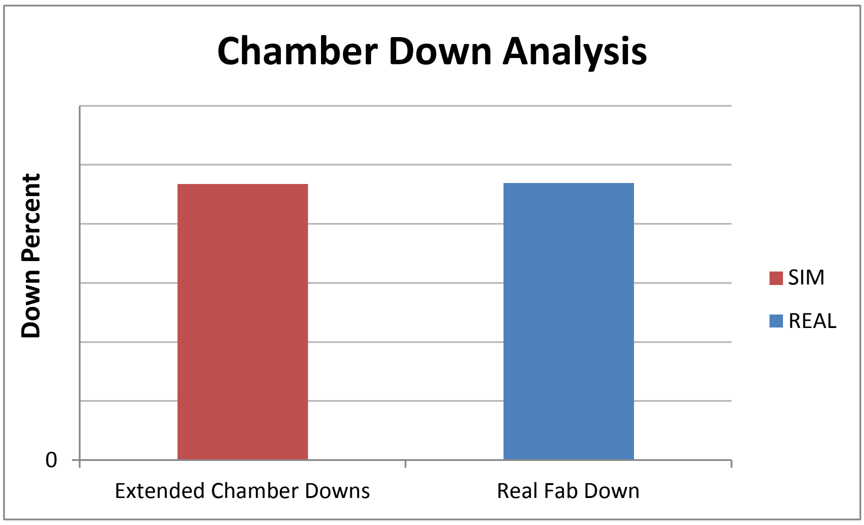


Figure 28: Extended chamber down approach

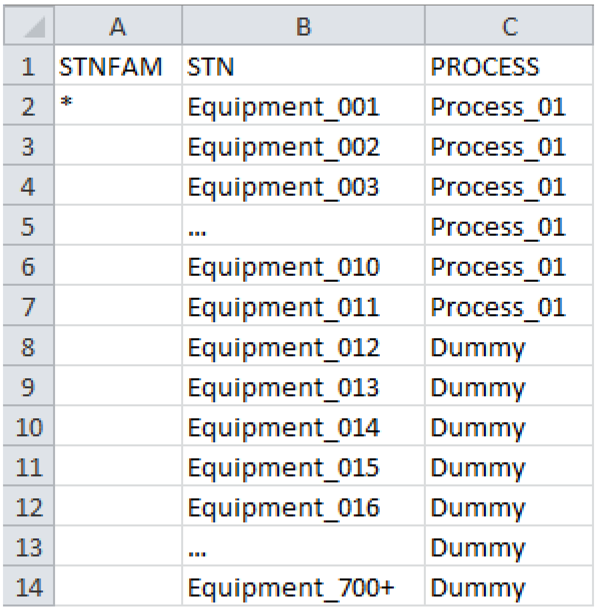
5.3.4 Equipment Grouping

For simulation modeling of a complex wafer fab, the problem of an appropriate equipment grouping exists. In practice, multiple different work center grouping criteria exist. The main criteria are the hardware configuration and the process definition. The hardware configuration grouping criteria is used for equipment maintenance. The process definition is used for capacity planning.

For simulation modeling the problem of an appropriate equipment grouping also exists. For simple models, without dedication, this is not a problem. A work center definition is used to define a set of parallel equipment, where the lot is able to run. The route step directly connects the used work centers. For complex dedication models, the concept of a process definition is useful. The route step defines which process is required. The equipment itself indicates which processes are able to run on the equipment. So an equipment grouping is not compulsory necessary but useful. The following simple experiment shows that the work center based equipment assignment definition is still important. It influences the simulation computation time.

A very simple but realistic model is used to show the effect. For the performance test only one route step is part of the model. For this single step 100000 lots are executed. The simulation execution time is measured. For this single step only 11 pieces of equipment are capable to process the lots. These 11 pieces of equipment are grouped into 3 different work centers in reality. The whole model contains more than 700 pieces of additional equipment. These pieces of equipment are not used for this particular step. They are required to compare the results of the simple model with the full fab model. For the full model there are also 700 pieces of equipment, which cannot be used for the particular step.

Figure 29 depicts the used process dedication for the equipment. The process “Process_01 is required by the executed step and the “Dummy” process is not used. Other model parameters are self-generated. The release rate is set to one lot per minute and the process time is set to 10 minutes. The used simulation software is ASAP.

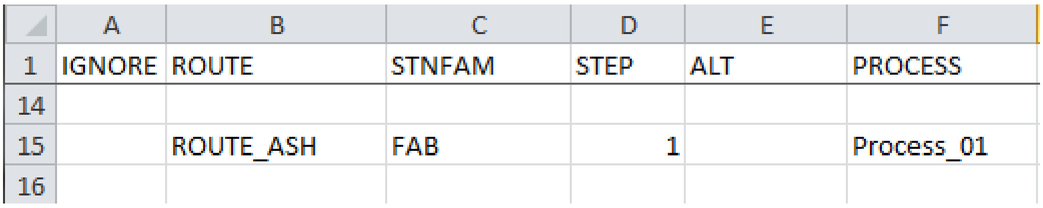


	A	B	C
1	STNFAM	STN	PROCESS
2	*	Equipment_001	Process_01
3		Equipment_002	Process_01
4		Equipment_003	Process_01
5		...	Process_01
6		Equipment_010	Process_01
7		Equipment_011	Process_01
8		Equipment_012	Dummy
9		Equipment_013	Dummy
10		Equipment_014	Dummy
11		Equipment_015	Dummy
12		Equipment_016	Dummy
13		...	Dummy
14		Equipment_700+	Dummy

Figure 29: Equipment dedication

The experimental design compares three different equipment grouping approaches. In scenario one, all pieces of equipment are in one group. In scenario two the work center grouping concept is used like in reality. In scenario three no equipment group is used. The station family value, marked with “*” in Figure 29, depends on the particular scenario.

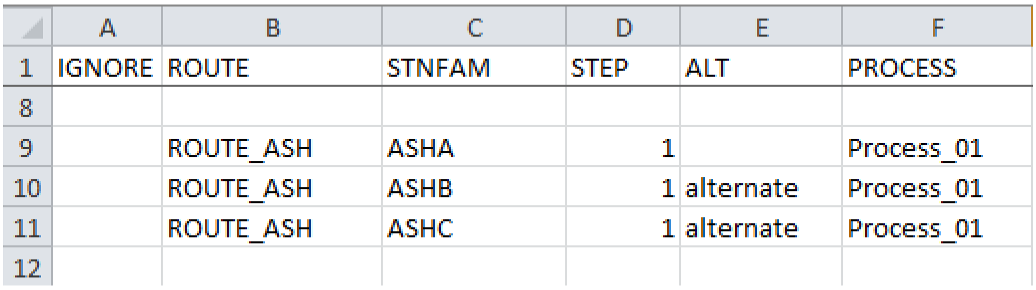
For scenario one, all pieces of equipment are available in one group. Figure 30 depicts the screenshot of the equipment grouping model in the route definition. The equipment group “FAB”, in the station family column, contains more than 700 pieces of equipment. Only a few pieces of equipment are able to execute “Process_01”. This process is required to execute a lot at step “1” for route “ROUTE_ASH”.



	A	B	C	D	E	F
1	IGNORE	ROUTE	STNFAM	STEP	ALT	PROCESS
14						
15		ROUTE_ASH	FAB	1		Process_01
16						

Figure 30: Scenario 1, Single route step, modeled with one work center for the full fab

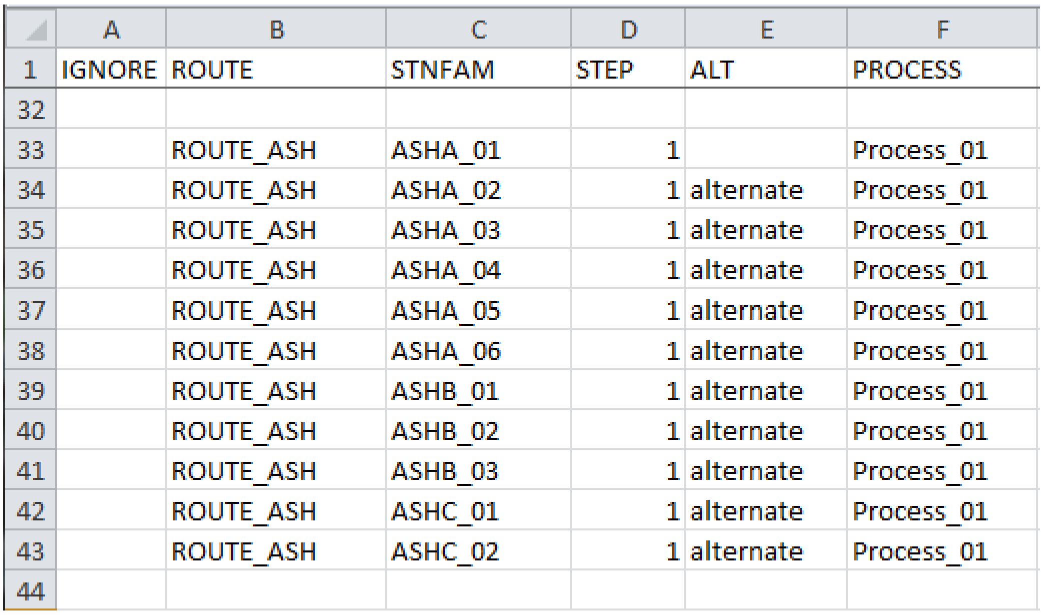
The second scenario consists of a work center definition from capacity planning, see Figure 31. All equipment, which is able to run “Process_01”, is distributed to three different work centers. Those work centers have been assigned as “alternate”, so a lot can be executed within all of those work centers.



	A	B	C	D	E	F
1	IGNORE	ROUTE	STNFAM	STEP	ALT	PROCESS
8						
9		ROUTE_ASH	ASHA	1		Process_01
10		ROUTE_ASH	ASHB	1	alternate	Process_01
11		ROUTE_ASH	ASHC	1	alternate	Process_01
12						

Figure 31: Scenario 2. Single route step, modeled with three alternate work centers

The third scenario is based on a grouping definition, where each work center has one specific piece of equipment assigned (Figure 32). Each piece of equipment, out of those work centers, is modeled as an "alternate" step. The process assignment will still work with the process definition which is used for comparability.



	A	B	C	D	E	F
1	IGNORE	ROUTE	STNFAM	STEP	ALT	PROCESS
32						
33		ROUTE_ASH	ASHA_01	1		Process_01
34		ROUTE_ASH	ASHA_02	1	alternate	Process_01
35		ROUTE_ASH	ASHA_03	1	alternate	Process_01
36		ROUTE_ASH	ASHA_04	1	alternate	Process_01
37		ROUTE_ASH	ASHA_05	1	alternate	Process_01
38		ROUTE_ASH	ASHA_06	1	alternate	Process_01
39		ROUTE_ASH	ASHB_01	1	alternate	Process_01
40		ROUTE_ASH	ASHB_02	1	alternate	Process_01
41		ROUTE_ASH	ASHB_03	1	alternate	Process_01
42		ROUTE_ASH	ASHC_01	1	alternate	Process_01
43		ROUTE_ASH	ASHC_02	1	alternate	Process_01
44						

Figure 32: Scenario 3. Single route step, each piece of equipment represent one work center

The experimental results are available in Table 17. For scenario one, the execution time for the simulation run is much longer than the run for the other two scenarios. The reason behind this is that all equipment of the whole fab is within this one group. For one single step only 11 pieces of equipment can handle the required process. But to assign a lot to the equipment, the algorithm will search in a list with more than 700 pieces of equipment. Within this list, only those 11 pieces of equipment are available, where the required process is available too. This is very inefficient and not necessary.

The second scenario with a proper work center definition shows significantly better results. The execution time of 45 second is much lower than the time from all other scenarios. The reason behind is that only a few work centers with only a few pieces of equipment have to be investigated if an appropriate equipment is able to handle the required process.

The third scenario, with one piece of equipment per work center, also provides acceptable results. The execution time is 52 seconds. The disadvantage of this approach is that the memory requirements are too high. For each step, all pieces of equipment, where the lot is able to run, have to be defined (Figure 32). Instead of one row for the fab approach, or three rows for the work center approach, eleven rows become necessary. For a real fab model with several routes, hundreds of steps and multiple pieces of available equipment per step, the data volume becomes enormous.

The conclusion is that, it is highly useful to deal with an appropriate equipment grouping concept for model size reduction and execution time reduction. This kind of strategies is very important because of the real time requirements of online simulation. For this example, a proper work center grouping shows the best results. The execution time is lower compared to the other scenarios. The required memory is also still acceptable. This approach is used for the short term simulation model.

Scenario	Grouping criterion	Simulation execution time
1	Whole fab	73 s
2	Work center	45 s
3	Equipment	52 s

Table 17: Performance results

5.3.5 Routes and Products

For route and also for product modeling it is necessary to define the adequate level of detail. Two basic options are available. The first option is to use real routes like they are used in the MES. The second option is to use a higher level of abstraction and combine similar routes to one abstract route, see Table 18. This option is feasible because several routes differ only in very few steps. The advantage of the second options is that it is useful to reduce data volume in the simulation model. The disadvantage is that a mapping to abstract route names becomes necessary.

Real Route	Abstracted Routes
LX4_001	LX4
LX4_007	
MC9_023	MC9
MC9_042	
MC9_044	

Table 18: Illustration for route abstraction

The decision has been made to use real routes names and real product names because in a complex manufacturing area a lot of information and data tables are linked to the real route names. So it is useful to use real route names in order to avoid a mapping between real route names and abstract route names for each table. Another reason to do so is that there are details getting lost due to abstraction.

5.3.6 Attribute Sampling

In semiconductor manufacturing lot attributes are used for operational control. Those control decisions are, for example, dispatch priority, equipment allocation, and sampling. Table 19 depicts an illustration of lot attributes. For short term simulation the idea is to increase the sampling modeling accuracy by using lot attribute information (Scholl et al. 2011). It is useful to consider lot attributes, to determine which lots enter an operation and which lots do not. The reason behind that is the attributes are already available at simulation start. For long-term simulation the exact lot value is not available yet. Therefore the common modeling way for sampling is to apply an average sampling rate for each step, derived from history.

Attribute No.	Type	Description	Values
01	Sampling	Enter the sampling step when the step number is the attribute value.	Step 389, 876
02	Sampling	Enter operation when the sampling class of the lot matches the criteria for an operation.	A, B, C
03	Sampling	Enter a particular sampling operation when attribute is set to yes.	Y, N
04	Priority	The priority classes of the lot effect the sorting in the dispatch list.	1,3,5
05	Priority	This flag for lot priority affect the sorting of the dispatch list.	Y, N
06	Equipment allocation	Indicate to which equipment the lot will be assigned.	Equipment_A_01, Equipment_A_03

Table 19: Lot Attribute Examples

Preconditions

To apply attribute based sampling, following preconditions apply.

- The lot attribute is related to sampling at certain operations.
- The lot attribute is available at simulation start and applies for the whole simulation time horizon.

The first condition has been checked by a questionnaire and the system documentation. Only those attributes are used which have a certain impact on the sampling decision. For the second condition a more detailed analysis is required. Only those attribute are useful, where the values are available at simulation start, and if these values affect the whole simulation forecast horizon. It means, the value will not be changed within the forecast horizon. Therefore it is necessary to determine the effective interval of the attribute. The effective interval of an attribute is the time between where the lot attribute is set (or changed) and the time where the attribute takes effect. The time when it takes effect is typically the enter operation time of the lot at the sampling operation.

If the effective interval is larger than the simulation horizon, it is feasible to determine the lot behavior at a sampling step with high precision. If the effective time is shorter than the forecast horizon, than only those attributes values are available at simulation start, which affect only the first days of the forecast period. For the last days of the forecast period, the correct attribute values are not yet available at simulation start. In this case a common sampling approach is practicable without considering lot attributes.

To determine if an attribute is suitable, a histogram of the effective interval is very useful. Figure 33 shows the histogram of the effective interval for attribute 01. The statistics collection refers to a representative period of one week. About 380 lots with the attribute 01 have an effective interval of less than one day. For most lots, about 420 lots in the diagram, the effective interval is slightly more than one day (between 24 and 48 hours). Only few lots have an effective interval of 7 days or more. The desired horizon of the simulation forecast is 7 days. So, attribute 01 is not used because the effective interval is less than 48 hours for most of the lots. It is too short compared to the simulation forecast horizon.

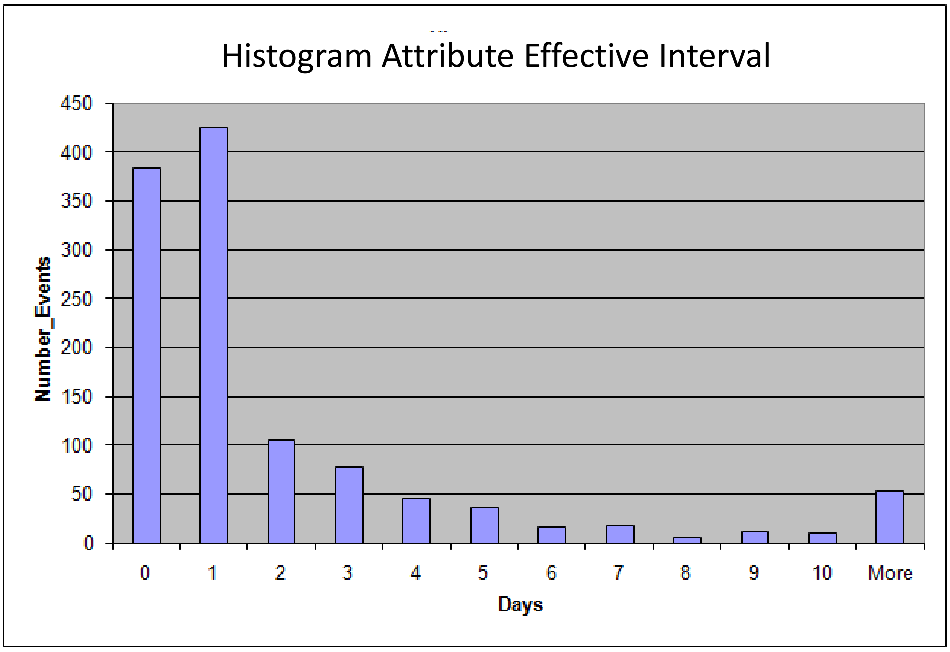


Figure 33: Effective interval histogram for attribute 01

In Figure 34 the histogram of the effective interval for attribute 02 and 03 has much more potential. Most lots have an effective interval of seven days or longer. The reason for this is that both attributes are already available at lot release in reality. Here the effective interval represents the time between lot release and the time when a lot reaches a sampling operation.

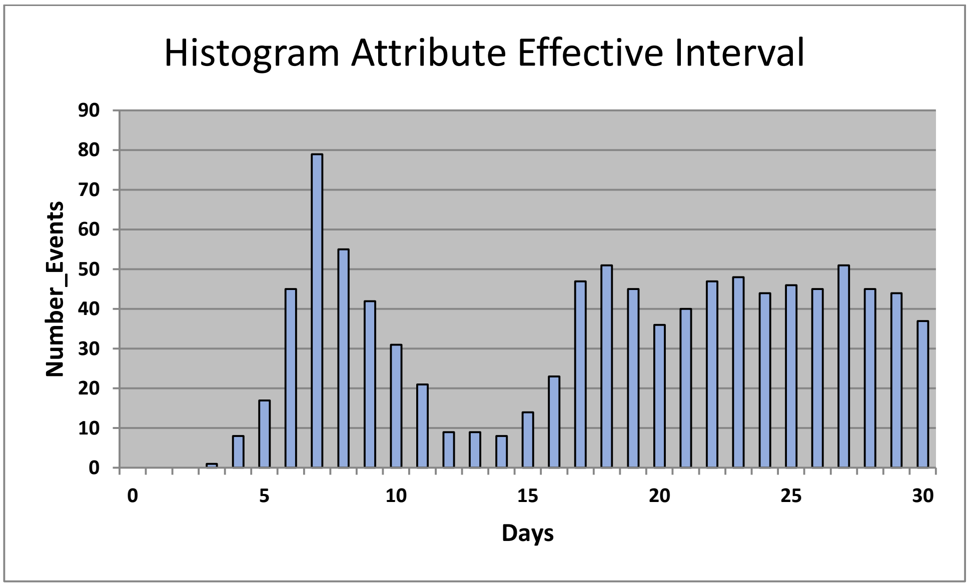


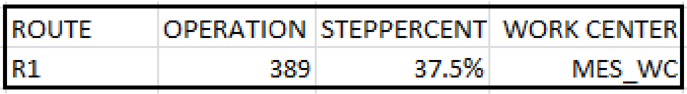
Figure 34: Effective interval histogram for attribute 02 and attribute 03

Implementation Steps

Following steps are required to implement attribute based sampling:

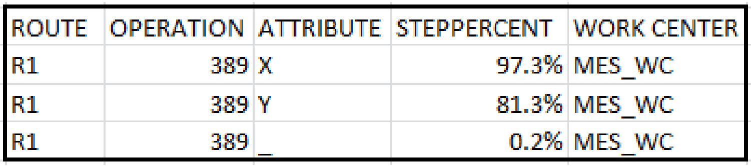
1. Identify sampling related lot attributes
2. Identify work center or route steps, where these attribute take effect
3. Check the effective interval of the lot attribute
4. Compute the attribute dependent step percent values by historical analysis
5. Extend the step sampling rates with attribute dependent step sampling rates
6. Initialize the fab with the attribute values for each lot

Step 1 and 2 are conducted through interviews with a specialist of the measurement department. The sampling related attributes and the affected route steps have been discussed. In addition to that, the impact of the lot attributes and the percentage of affected lots was also part of the discussion.
In step 3, it is necessary to analyze if the effective interval of the lot attributes is larger than the simulation horizon, see the description in the previous section.
In step 4 the system generates the sampling rate for the lot attributes 02 and 03, using the historical lot trace of a 60 days period. The computed sampling rate value depends on the route, the operation, and the attribute value. The system executes the historical analysis weekly, to keep the sampling rate values up to date.
Step 5 extends the simulation model with the attribute based sampling approach. The attribute sampling extends the common step sampling rate of one step to an attribute dependent step sampling rate. Table 20 shows an example of a regular sampling rate, which depends on the current route step. The operation 389 in route R1 has a step percentage of 37.5%. With the extension in Table 21 the lot attribute value “X”, “Y”, and “_“ defines a custom sampling rate “STEPPERCENT” for the same route step. Lots with attribute “X” execute this step with a probability of 97%. Lots with attribute “_” execute the step with a chance of less than 0.2%.



ROUTE	OPERATION	STEPPERCENT	WORK CENTER
R1	389	37.5%	MES_WC

Table 20: Sampling modeling without attributes



ROUTE	OPERATION	ATTRIBUTE	STEPPERCENT	WORK CENTER
R1	389	X	97.3%	MES_WC
R1	389	Y	81.3%	MES_WC
R1	389	_	0.2%	MES_WC

Table 21: Attribute dependent sampling rates

The attribute based sampling approach is implemented as a custom extension in ASAP. When a lot arrives at an operation, a table look-up is performed to check, if the route-operation combination exists in this attribute extension table. If an entry is available, the sampling rate applies, depending on the attribute value. If no entry is available, the step percent of the route-operation pair is used.
In Step 6 it is necessary to initialize the simulation model with lot attributes. For all WIP lots and all future lot releases, the lot attribute values are available in the model. It is either available in the lot release data source or it is derived through the product number.

Results

To compare the modeling accuracy of both approaches the work center wafer arrival performance parameter is used. The sampling modeling approach affects the arrivals directly.
Therefore it is useful to compare the arrival rates for evaluation of the sampling modeling approach. Figure 35 depicts the wafer arrivals for one work center. It can be seen that the arrival modeling without attributes is not sufficient for a work center. The number of arriving wafers deviates too much from reality. By including the attribute information the difference between simulation and reality improves a lot (Figure 36).

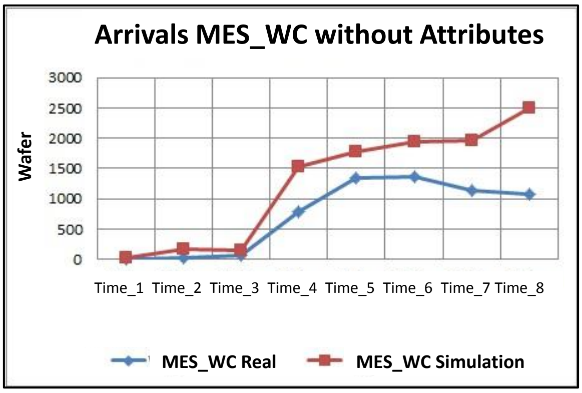


Figure 35: Wafer arrivals without attributes

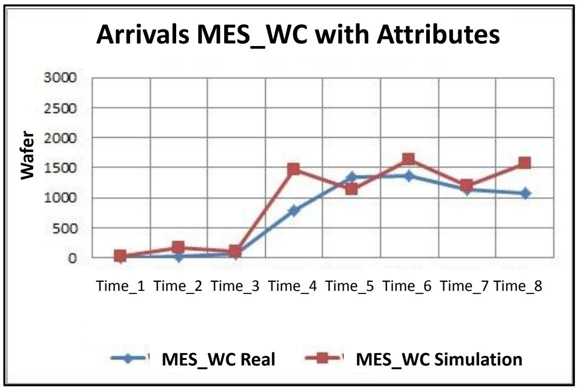


Figure 36: Wafer arrivals with attributes

A lot of effort has been spent on analysis, data access, and model extension. The conclusion is that the consideration of sampling related lot attributes is useful to improve work center modeling accuracy. Unfortunately for many other lot attributes, the effective interval is too short, compared to the simulation time horizon. If the time when the lot attributes are set and the time when they take effect, is too short, this type of lot attribute is not valid for the end of the simulation forecast time horizon.

5.3.7 Dedication

To implement a proper dedication model, the first step is to investigate which dedication elements exist in the fab. An investigation has been carried out, which includes department interviews, the analysis of dedication software tools, the analysis of dedication data sources, and the review of the source code of the dedication implementation.

One challenge of this investigation is to understand where the blocking system takes effect. Another challenge is to figure out the reason, why a lot is blocked for a particular process. The next challenge is to find a proper abstraction level to describe and model these systems. For many dedication elements the concept is very similar, but the implementation, the data sources, and dependencies are very different.

The following results are the outcome of the dedication investigation. It turns out, that many local and heterogeneous dedication systems exist. All together more than 20 blocking elements have been analyzed but not all of them have dedication reasons. Table 22 contains an overview of different dedication elements and their location. Some dedication elements apply for the whole fab, others affect all work centers from a production area, and elements also affect only a single work center within the production area.

Blocking Elements	FAB	WETBENCH _01	CLUSTERTOOL_01	METROLOGY_02	CLUSTERTOOL_05	LITHOGRAPHY_01	METROLOGY _01	IMPLANTATION_09	….
Standard process dedication	X								
Location				X					
Backup tool activation		X	X	X	X	X		X	
Lot blocked, Blacklist	X								
Litho horizontal dedication blocking						X			
Litho vertical dedication blocking						X			
Chamber up blocking			X		X				
Test program assignment				X			X		
Wet bench chamber blocking		X							
Implant process blocking								X	
…									

Table 22: Work center dedication blocking matrix

To provide an impression of the purpose and the complexity of dedication elements, the following explanation part describes some dedication elements from the list above. Table 23 provides an overview of the dependencies of the dedication elements as well. It is necessary to understand how a single dedication element works, before the overview of the dependencies becomes available.

The **standard process dedication** indicates that not all equipment of a single work center can handle similar processes. The reasons are resource and cost saving aspects. The decision whether the lot can be processed depends on the equipment name and the process name.
Another element is the **location blocking** feature. It is not allowed, to process two subsequent steps of a lot on two different pieces of equipment. The reason is cycle time reduction, if the equipment is located too far away from each other. The dependency is the location of the source equipment and the location of the target equipment.
The **horizontal lithography dedication blocking** indicates that not all products are able to run on any equipment, even if the process is available for those tools. The dedication becomes product specific. So the dedication element depends on the equipment, the process, and the product as well.
The reason for **vertical dedication for lithography** tools are process constraints. The same equipment, which has already processed a previous lithography layer of the lot, needs to be used for the next lithography layer of the lot. The dedicated equipment depends on the predefined operation and the equipment, which processed the lot earlier.
The **chamber up blocking mechanism** is mainly used for cluster tools with multiple chambers. Due to a chamber down not all processes are able to run on the equipment any more. The chamber down blocking mechanism defines which processes are still able to run, and which are not.

Dedication element	1st Key	2nd Key	3rd Key	4th Key	5th Key
Standard process dedication	Equipment	Process			
Location	Location of the source equipment	Location of the target equipment			
Backup tool activation	Equipment	Process	Priority		
Lot blocked/ Blacklist	Lot	(Operation)			
Litho horizontal dedication blocking	Equipment	Route	Process		
Litho vertical dedication blocking	Lot	Operation	Equipment		
Chamber up blocking	Process	Chamber up status			
Test program assignment	Equipment	Operation	Process parameter_A		
WET chamber blocking	Equipment	Product	Operation	Process	
IMP blocking	Equipment	Process	Process parameter_A	Process parameter_B	Process parameter_C
…					

Table 23: Dedication dependencies

A conclusion from this investigation is that so many different local dedication elements exist, so that the task to implement all in the fab model is too big. The benefit to implement a single dedication element is not so large because it only directly affects a small part of the fab model. The final decision is to use the standard process dedication and only a few custom dedication elements, which have high impact on fab performance. For the standard process dedication the process name is used in the model, to define which equipment is capable to run a particular process within a work center.

For lithography a dedication element has been implemented to increase the model accuracy, because lithography is the default bottleneck. Therefore the capacity modeling for lithography needs to be very accurate. The horizontal dedication feature has been added to the fab model. A very generic approach is used, to keep the dedication feature extendable to other work areas beside lithography. The analysis above shows, that for lithography, the key is defined by equipment, route, part, and process. The dedication approach extends the model with the product dependency, to reflect lithography dedication blocking. In Table 24 an example is depicted for such a combination which is blocked. During task selection the simulation tool checks whether the combination is blocked or not.

STN	ROUTE	PART	STNCERT
EQ_02	Route_01	Part_01	CertA

Table 24: Blocked process combination

Many test cases have been run to check several dedication blocking combinations. The result of the simple test case for the combination above is depicted in Figure 37 and in Figure 38. Both lots from product "Part_01" on "Route_01" need to execute three subsequent steps with Process "CertA". The work center contains two stations. If the dedication is available the lots run parallel on these stations. If the dedication combination from Table 24 is blocked, than both lots are not allowed to run on equipment "EQ_02".

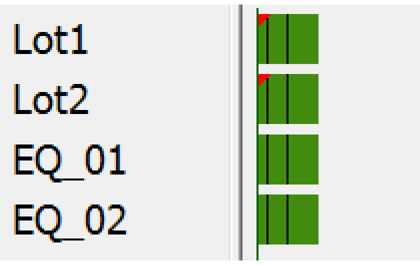


Figure 37: Blocking disabled

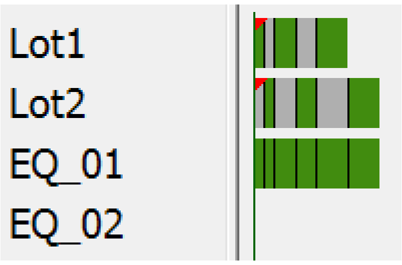


Figure 38: Blocking enabled

To enhance the usability, a concept of wildcards "*" has been implemented for dedication blocking. Another concept is to define a positive and negative list to define which combination is allowed or not. If only a few combinations are blocked, it is not useful to list all combinations which are not blocked. This concept reduces the data volume and the computation time.

5.3.8 Dispatching Rules

Dispatching rule modeling is a major part of the wafer fab model. In a mature wafer fab, functionality has been increased over the past years to improve the fab performance. The complexity of the dispatching rules becomes large. Multiple ways exist to model the dispatching rules.

One option is to reuse such components of the control mechanism. The objective is to add a lot of functionality to the model without much effort. In literature interesting approaches also exist to use the same control logic in the real fab and in simulation (Smith et al. 1994). The purpose is to reduce the effort for the development of the control logic in simulation and reality. Mönch et al. (2003) also present an approach where external control logic interacts with the simulation model. To reuse such components in a mature wafer fab, several disadvantages arise. The following section elaborates on these disadvantages in detail:

- High complexity due to a high number of heterogeneous control mechanisms
- Data for future control decisions does not exist yet
- Reduced system performance

In a mature wafer fab, a **high number of heterogeneous control mechanisms** exist. Examples are the lot release mechanism, the reticle handling, the sampling controller, the heterogeneous PM plan, and numerous local dispatching rules. Mönch et al. (2003) also pointed out, that not all information is available in the state-of-the-art core MES components.
It is a problem because the integration effort to use those mechanisms in simulation is high. Every control mechanism follows a different standard. Several control decisions, for example, test wafer release, are executed manually.

Another issue is that **future data for future control decisions does not exist yet**. For example a dispatching decision in reality uses the current lot state, including all lot attributes. The same controller needs all that information for a future decision in simulation, too. Unfortunately most of the lot attributes are not yet available. It becomes necessary to emulate future data like lot attributes (Smith et al. 1994). The creation of an emulation engine for future data increases the effort dramatically. In other words, it reduces the benefit of sharing the same controller in reality and in simulation. Another disadvantage is that the accuracy of control decisions, based on emulated data is reduced.

Another argument is the **reduced system performance**. To discuss the effect on system performance, it is necessary to discuss the way of connecting simulation and fab controller. Basically two approaches exist to operate with the same fab controller:

- The first option is to use a communication link between fab controller and simulation model. The controller of the wafer fab handles the predefined decisions of the simulation model. An interface becomes necessary to use the existing control mechanism. This leads to a high load on the MES side, which is not acceptable. The time consuming data transfer between the simulation and the fab controller also reduces the simulation speed significantly.
- The second approach imports the fab controller into the simulation model. This concept has the advantage that there is no time loss due to the import/export of data. Unfortunately the control decision is as complex as in the real wafer fab. By adapting the same control mechanism 1:1, an abstraction is missing. This also leads to reduced speed for the short term simulation. It is a potential violation of the real time requirement.

The conclusion is that in simulation a controller is used, which is independent from the real fab. Due to the high complexity of the real wafer fab, this controller is an abstraction of the real world. From the implementation point of view, the controller is represented by the simulation engine, self-generated simulation extensions, and the implemented dispatching rules.

Regarding the dispatching rule itself, the requirement is to mimic the behavior of the dispatching rules closely to reality. The task is to implement the dispatch rule on a proper abstraction level, to achieve good simulation results within a limited time horizon. The dispatch rule has been implemented as an ASAP extension. It is a combination of dispatching rules where standard and custom rules are used. Due to the non-disclosure agreement, the details which dispatch rule is used are not part of this thesis.

5.3.9 Kanban Modeling

In the wafer fab Kanban becomes necessary because of timing constraints between different processes. For those time bound sequences, the Kanban WIP limit reduces the number of lots to guarantee that the lots are processed within their time limits. To realize Kanban, the number of lots between process start of operations n and process finish of operation n+1 has to be limited to a certain WIP level, see Figure 39.

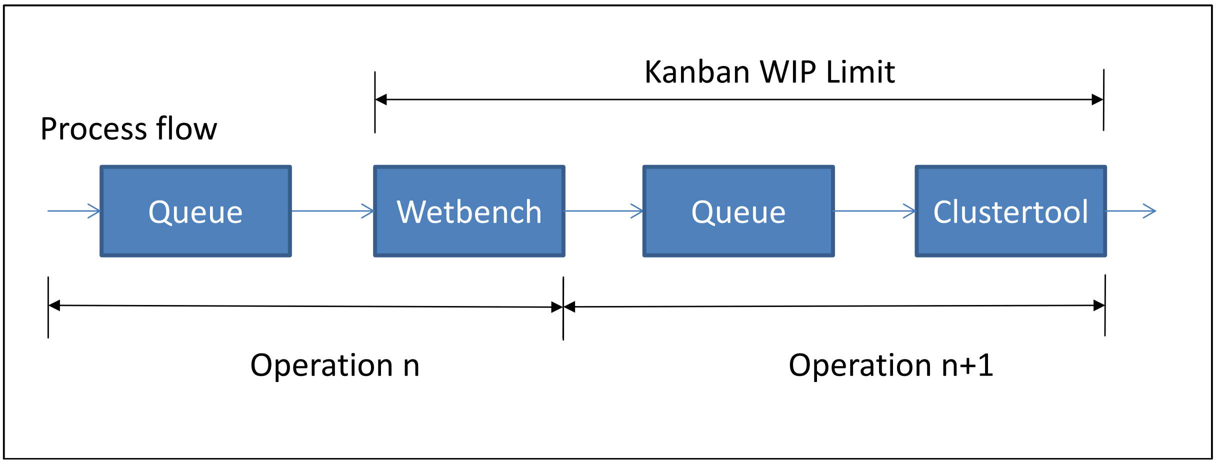


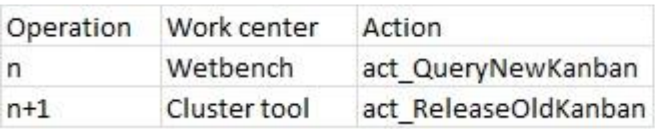
Figure 39: Kanban modeling

Kanban concept
In the wafer fab many local Kanban systems exist for several work centers. Each Kanban system differs in its configuration:

- Number of steps within the Kanban system
- Lot limit of the whole Kanban system
- Dependency of the Kanban lot limit on different products
- Dependency of the Kanban lot limit on equipment downs

The objective is to decide if the Kanban is necessary to improve the forecast. It is necessary to compare a forecast with and without the Kanban implementation. Therefore the Kanban has been implemented. In a second stage it is also feasible to think about the optimization of the Kanban itself. The limit of the Kanban lots changes dynamically, depending on equipment downs, future lot arrivals, and the product of the lots. At the moment the Kanban lot limit is conservative. It means that a throughput loss is acceptable to avoid a costly violation of a time bounded sequence. The benefit of the simulation based optimization approach is that the throughput loss is reduced without increasing the risk of a cycle time loss.

Kanban implementation
In order to realize the Kanban feature, a Kanban custom extension is implemented in ASAP. For operation n, the action "act_QueryNewKanban" will be executed before resource allocation (Table 25). If no more Kanban tokens are available the lot has to wait. When a token becomes available it continues processing. For operation n+1 the action "act_ReleaseOldKanban" will be executed after processing. Therefore the Kanban token is released. The Kanban is implemented to ensure the process time constraint between a wet bench process and a cluster tool process.



Operation	Work center	Action
n	Wetbench	act_QueryNewKanban
n+1	Cluster tool	act_ReleaseOldKanban

Table 25: Custom actions for Kanban operations

Kanban Results
The results of the implemented Kanban system are available in Figure 40 up to Figure 43. The figures show the WIP results with and without Kanban for the wet bench and the cluster tool.

Without the Kanban implementation, the WIP at the first Kanban operation, the wet bench tool, is too low. The WIP for the last operation, the cluster tool, is too high. The forecast results deviate considerably from reality. After the Kanban has been added to the simulation, the wet bench and the cluster tool WIP are much better. The lots queue up at the wet bench work center and not at the cluster tool any more.

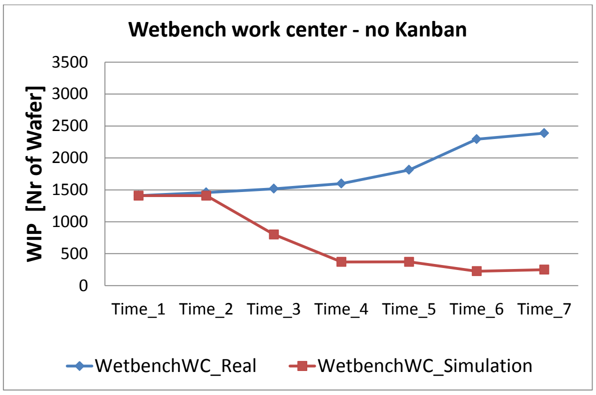


Figure 40: Wet bench without Kanban

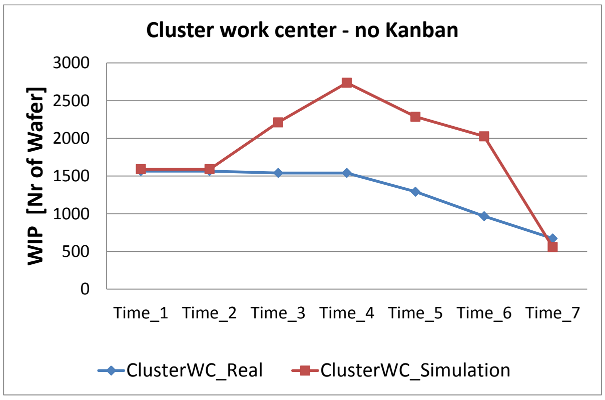


Figure 41: Cluster work center without Kanban

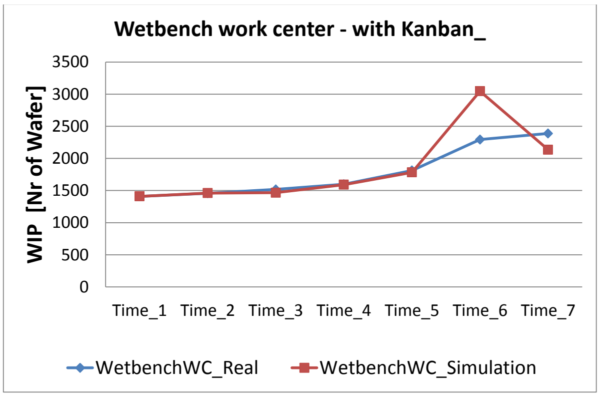


Figure 42: Wet bench with Kanban

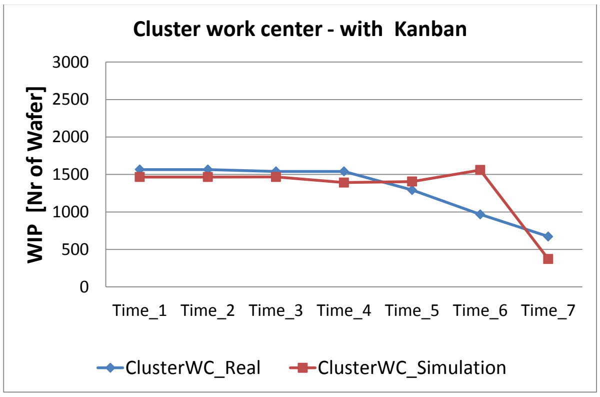


Figure 43: Cluster work center with Kanban

Kanban Conclusion

The conclusion is that Kanban is highly relevant for forecasting. With the Kanban implementation in the simulation model, the WIP is assigned to the correct work center, which reflects reality best. For results interpretation, it is necessary to consider the Kanban limit. There is the danger to use the forecast results of a Kanban system without the internal knowledge of that Kanban system. The reason is that the WIP of a potential bottleneck work center is limited to the Kanban threshold. So the WIP for the previous work center within the Kanban increases. In this case it is not correct do derive that this previous work center has performance problems. It is necessary to consider the Kanban constraints, to figure out where the root cause of a performance problem lies.

5.3.10 Send-Ahead Modeling

To model the send-ahead wafer modeling feature, it is necessary to derive the split rate, the production step for splitting, and the production step for merging, see Table 26. This information is derived from historical data analysis. The feature itself is implemented as an ASAP custom extension. The split name and combine name are used to refer to the related operations.

ROUTE	STEP	SPLITNAME	SPLITSIZE	SPLITPERCENT	COMBNAME
Route_01	001				
Route_02	002	Route01_Step002	1	0.35	
Route_03	003				
Route_04	004				Route01_Step002
Route_05	005				

Table 26: Send-ahead modeling

The underlying model data captures the following fab behavior, like seen in the example in Figure 44. The lot is being processed as normal for operation 001. In the beginning of operation 002 the child lot with one wafer splits up from the mother lot. It is processed in operation 002 and 003, while the mother lot is waiting. Before the child lot reaches operation 004, the mother lot needs to process operation 002, 003. At the beginning of operation 004, the mother and child lot merge. They are processed as normal for operation 004 and 005.

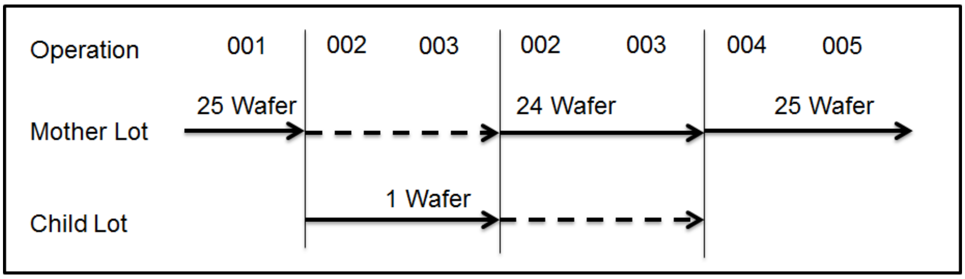


Figure 44: Send-ahead mechanism example

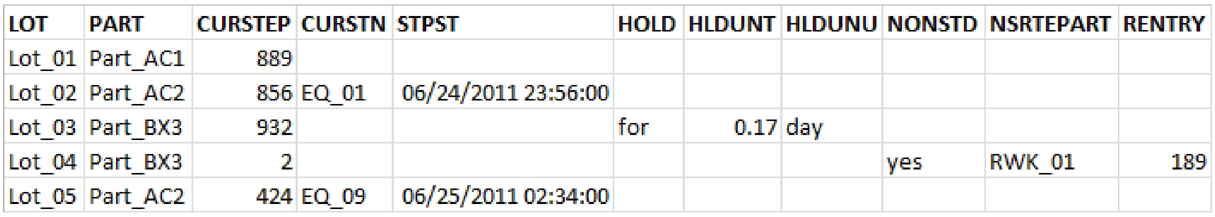
5.4 Model Initialization

The short term simulation model highly depends on the initial state (Reijers and Aalst 1999). Therefore the model initialization section is a major part of online simulation. This section describes the modeling elements which are related to model initialization. The description includes the concept, the implementation, and the analysis of the simulation results.

5.4.1 Lot Initialization

Lot initialization has two aspects, the modeling of future lot releases and the current WIP modeling. First the initialization of WIP lots will be discussed.

The key challenges to warm start the online simulation model is to initialize the WIP lots properly. The basic approach initializes each individual lot. So, for each lot in the fab, the current operation number is available. It defines at which operation the lot starts its route into the model. Table 27 illustrates that for each lot the current lot operation “CURSTEP” will be assigned. Due to non-disclosure agreement, Table 27 contains anonymous data.



LOT	PART	CURSTEP	CURSTN	STPST	HOLD	HLDUNT	HLDUNU	NONSTD	NSRTEPART	RENTRY
Lot_01	Part_AC1	889								
Lot_02	Part_AC2	856	EQ_01	06/24/2011 23:56:00						
Lot_03	Part_BX3	932			for	0.17	day			
Lot_04	Part_BX3	2						yes	RWK_01	189
Lot_05	Part_AC2	424	EQ_09	06/25/2011 02:34:00						

Table 27: Initial lots with properties

If only the route step is assigned, every lot starts in the queue. To further enhance the model accuracy, it is useful to define which lots are in the queue and which lots are in process. For the processing lots it is furthermore important to define how long they are already in the process. The column "CURSTN" and "STPST" in Table 27 define the current equipment of the lot and the starting time of the process. If the field is empty, the lot is in the queue. If the current station and the step start time of the lot are defined, the lot is in process. Further lot properties, which are relevant for lot initialization, need to be defined as well. Examples are hold and rework. For hold, the column "HOLD", "HLDUNT", and "HLDUNU" define for how long the lot is on hold. For rework, the column "NONSTD", "NSRTEPART", and "RENTRY" defines the rework route and the reentry point to reach the original route.

Another aspect of a proper lot initialization is the initialization of future lot releases. Common simulation approaches use a product dependent lot release distribution to model future lot releases. For short term simulation a distribution is not necessary. Most information is already available for each lot which will be released. Therefore it is feasible to initialize the model with real lot instances. Table 28 depicts the lot release file. It contains the real lot name, the product (part) of the lot, the starting date, and additional attributes like the number of wafers (pieces). With this approach, it is possible to assign different lot attributes even if the product number is the same. The lot release plan for the next week(s) is used as a data source.

LOT	PART	PIECES	PRIOR	START
Lot_01	116	5	5	06/05/2012 06:00:00
Lot_02	552	12	0	06/05/2012 06:00:00
Lot_03	905	20	0	06/05/2012 06:30:00

Table 28: Future lot release

As it can be seen in Figure 45, the lot release plan which is used in the simulation model reflects the real future lot releases. Slight changes apply if the lot release plan changes between simulation start and the lot release event.

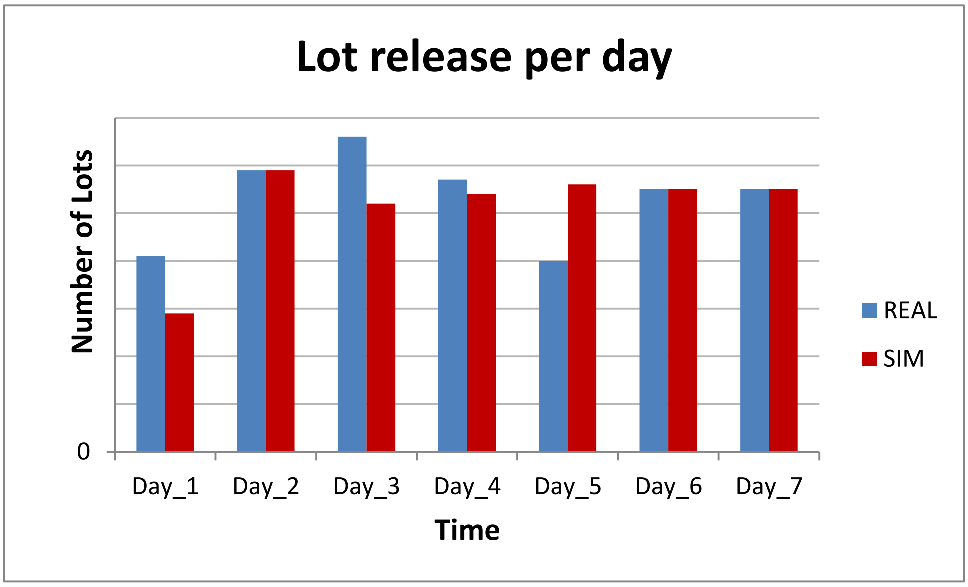


Figure 45: Lot release in reality and simulation

5.4.2 Sampling at Simulation Start

For the WIP lots initialization, it necessary to check if the work center WIP at simulation start is correct. Therefore Figure 46 shows the initial work center WIP for five work centers. The observation is that the initial WIP is wrong, especially for WC_01 and WC_02. An investigation figures out, that the error reason was caused by the sampling rate calculation, which has been applied for initial lots. This is a critical modeling error. In reality a WIP lot at an operation already executes the sampling calculation. If a lot skips an operation, it is assigned to this operation. So in simulation it is not necessary to execute sampling for WIP lots again. For work centers WC_01 and WC_02 the sampling rate has a significant effect on the initial WIP results.

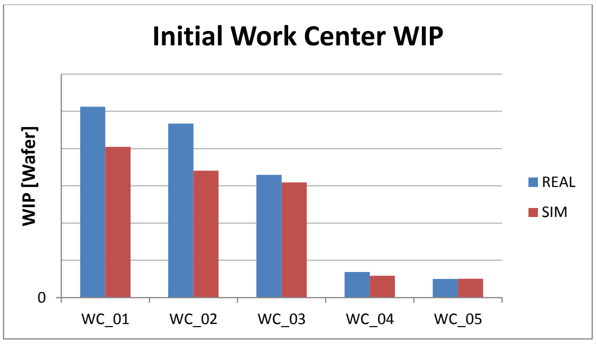


Figure 46: Initial WIP with sampling.

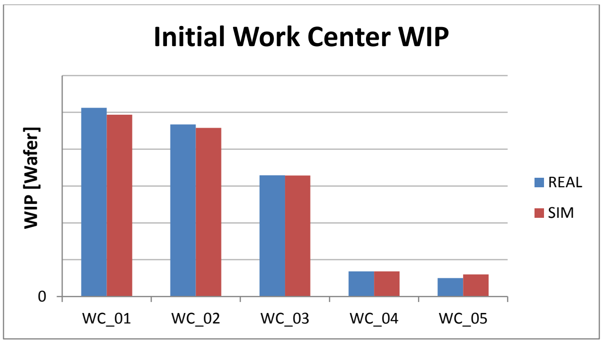


Figure 47: Initial WIP without sampling.

For testing purposes the sampling rate feature has been removed in the model. After that change, the simulation results reflect reality (Figure 47). Minor differences between simulation and reality still exist in the results. This is the common noise in a full fab simulation model. Typical reasons are for example data inconsistencies for development lots, whereby the current operation is not available. In this case the next available operation is used. So these lots are assigned to a different work center.

To deploy a permanent solution, it is necessary to avoid the sampling rate calculation for WIP lots. The solution is to use a separate actions sequence for initial lots. It is necessary to modify and separate the initial action list of the ASAP simulation. Depending on the lot state, the queue lot and the processing lot action list is used (Table 29). The default action list applies the sampling and the holds computation before the resource allocation. It executes the rework computation after leaving the process. For processing lots, it is not required to execute the sampling and hold computation. In reality the lot is already in the process. In simulation it is necessary to allocate the current station at the beginning and to apply the rework computation afterwards. For queue lots, it is not necessary to apply sampling in simulation. The rest of the action list is similar to the default action list. The lot may enter a hold state during queue time. It needs to enter the process by allocating resources. It also needs to execute the rework computation afterwards.

Default lot action list	WIP lot action list for processing lots
• Sampling • Hold • Execute standard resource allocation • Rework	• Execute current station resource allocation • Rework
	WIP lot action list for queue lots
	• Hold • Execute standard resource allocation • Rework

Table 29: Action list for WIP lots

5.4.3 Rework Initialization

The initialization of rework (Section 5.4.1) has been tested and it works as expected. The WIP lot on a non-standard route executes the non-standard route until it reaches the last route step. Afterwards, the lots reenter the original routing at the predefined reentry point.

5.4.4 Hold Initialization

Hold initialization is another aspect to further specify the modeling accuracy. Due to hold initialization, every hold lot in reality is on hold in simulation, too. The question is how to release the hold lots. A very general approach is to compute the average hold duration and use this value for initial hold duration of hold lots. The result is that all hold lots are released at the same time, see Figure 48.

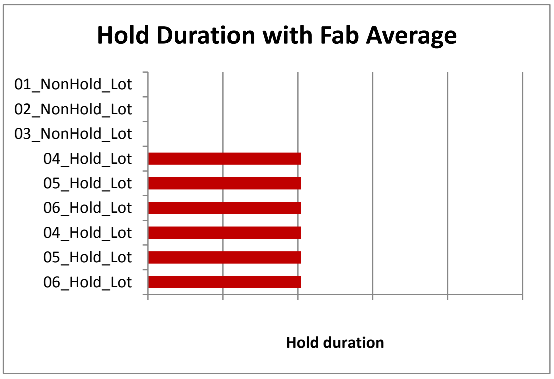


Figure 48: Average hold duration

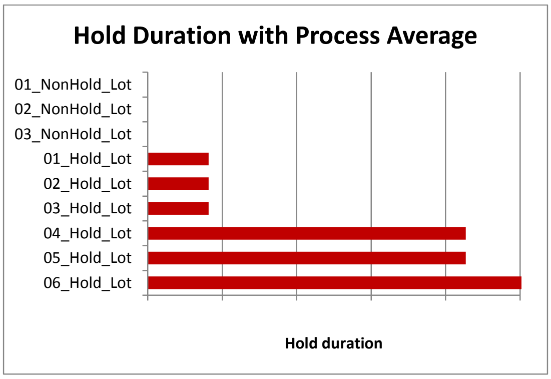


Figure 49: Average hold duration per process

To reduce the effect that all initial hold lots are released at the same point in time, several solutions are applicable. One solution is to use a distribution to determine the length of the hold lots. Another solution is to break down the hold duration into further categories. For online simulation, the second solution applies. The process category is used for the average hold computation.

This approach is used because the data for the process dependent average hold duration is already available. The results are depicted in Figure 49. For different process lots, the hold release time is different. Only for some lots the hold release time is still the same. Several hold lots exist within the same process. From online simulation perspective this approach is still sufficient to resolve the problem that all hold lots are released at the same time. The reason is that hold is not as important as other modeling features. Hold is not throughput relevant for the equipment. The average hold delay is a lot less compared to the transportation delay. The average hold delay is only about 10% compared to the average transportation delay. In the fab less than 2% of the lots are in hold state. So the percentage is very low. Hold is highly unpredictable as well. It is not possible to determine when a lot changes its hold state in the future.

5.4.5 Batching Initialization

This section analyzes the batch initialization. For a batch process, multiple lots start the process, execute the process, and finish the process together. To initialize a lot for a non-batching process, it is necessary to assign the current station and the start time of the running process. To initialize a couple of lots for a batch process, it is also necessary to assign the lots to the same batch. In reality all lots are using the same batch ID. For simulation the initial batch creation is specific to the simulation tool. In ASAP a batch ID is not necessary. The lots form a batch, if the current station is a batching tool and the step start time is the same.

In simulation a critical error occurs due to batch initialization if the start time is not the same. The following example helps to illustrate this behavior. As seen in Figure 50, the expectation is that batch “n”, with four lots, executes the remaining process time on the batch equipment. When the process is finished, the equipment releases the lots. The next batch “n+1” starts on the equipment. In reality the batch lots do not have the same starting time. It is slightly different. The difference is only a few seconds. Four different batches have been created in simulation. Each batch contains a single lot. The effect is that not all lots finish at the point in time (Figure 51). The effect is that many different process finish events are available in simulation. After each finish event, the simulation schedule has to run a new batch with four lots, which will also exceed the station capacity.

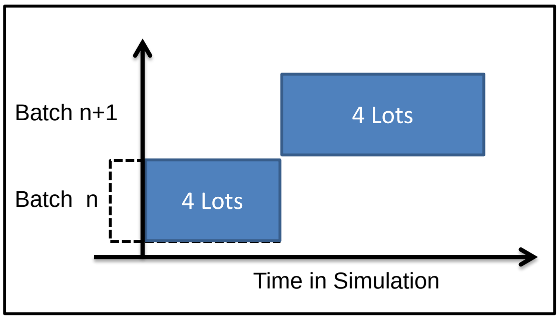


Figure 50: Correct batch initialization

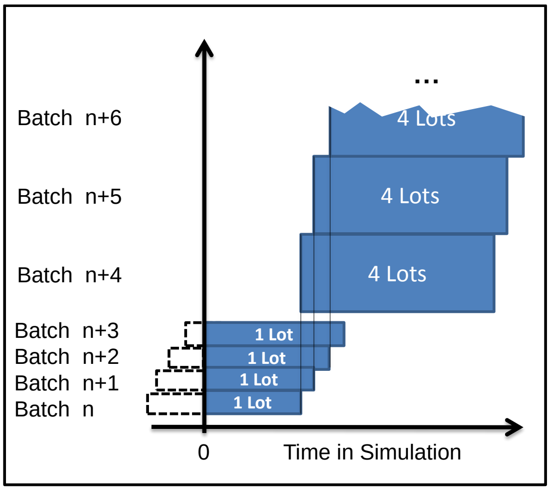


Figure 51: Incorrect batch initialization

To solve this problem for simulation, the step start time stamp needs to be the same. Instead of using the time stamp of the “move in” event, the time stamp of the “process start” event is used for all lots. This time stamp is more relevant for batching because the lots for one batch are able to come in at a different points in time. In the equipment they all start processing at the same time. A second control instance has been applied, to avoid incorrect model behavior due to bad data quality. If the start time is slightly different for batch lots, the step start time will be changed. This control mechanism enforces the same time stamp for batch lots. The solutions have been applied. The initial batch creation is corrected and meets the expectations.

5.4.6 Equipment Downtime Initialization

To initialize the fab model with the current fab state, it is also necessary to initialize the equipment with their current equipment states. For simulation start, the expectation is that the equipment throughput modeling is highly accurate, as long as the equipment in the model has the same state as in the fab. A proper initialization is also required for the average fab down. Without a proper initialization, the downtime modeling requires a warm-up period to take effect. Such a warm-up period is not acceptable for short term simulation. A very simple scenario illustrates this behavior. Assume that the MTTF and MTTR are defined as depicted in Figure 52. For all equipment in the fab, the MTTF is assigned to 90 hours. The MTTR is 10 hours. An exponential distribution is used. These values for MTTR and MTTF are self-generated values.

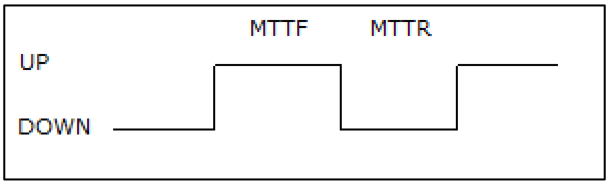


Figure 52: MTTF, MTTR definition in ASAP

According the definition in ASAP, the average downtime is computed according to the following formula.

$$AverageDowntime = \frac{MTTR}{MTTR + MTTF}$$

So for this self-generated example, the average downtime is 10%. The scenario runs for a period of 7 days simulated time. The most interesting part is the warm-up period for the downtime statistics. Figure 53 depicts the effect of the downtime distribution on the warm-up period. On the first day, the down percentage increases until it reaches the average of 10%. For rest of the 7 days period, it is alternating between 8% and 12%. In Figure 54, the warm-up period for initial downtime becomes evident. The figure shows the down percentage for the first day only. It takes about 9 hours to reach the average downtime in the fab. This behavior is crucial for short term simulation. The conclusion is that a proper downtime initialization is necessary.

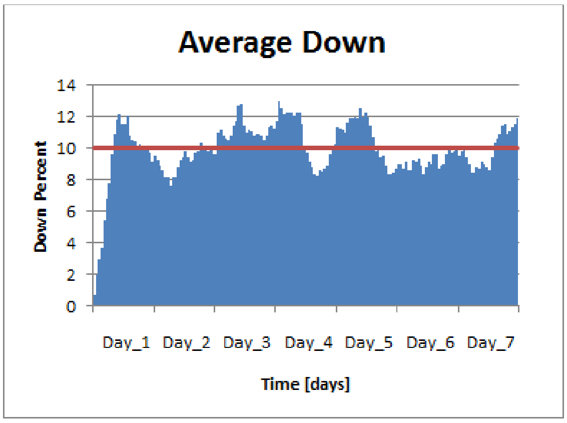


Figure 53: Average down

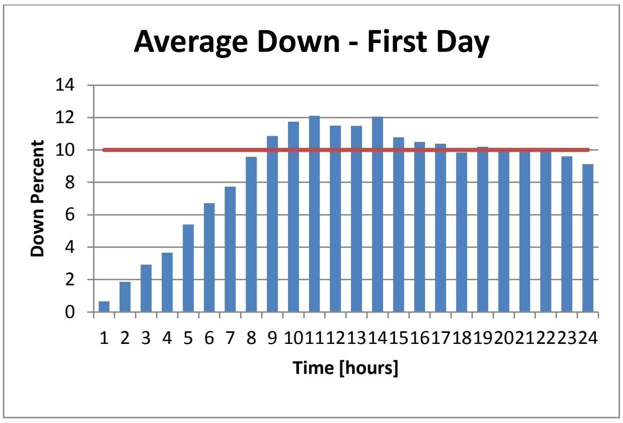
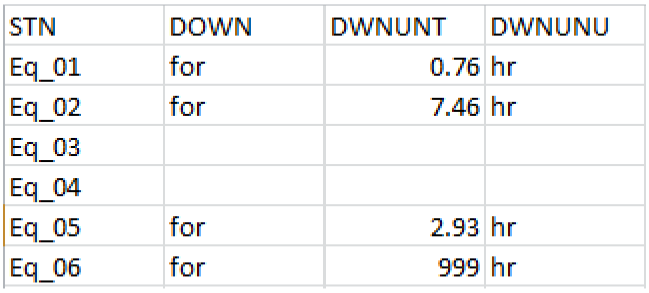


Figure 54: Average down on the first day

The solution is to initialize all equipment in the fab with their current up or down state. It is further necessary, to assign the estimated length of the current downtime. In the simulation model, see Table 30, the equipment state and the remaining equipment downtime are defined at simulation start. Equipment “Eq_01”, “Eq_02”, “Eq_05”, and “Eq_06” are in a down-state, while “Eq_03” and “Eq_04” are in an up-state at simulation start. For most equipment the downtime is a few hours. For “Eq_06” a user defined value is used. The equipment is down for the whole simulation period.



STN	DOWN	DWNUNT	DWNUNU
Eq_01	for	0.76	hr
Eq_02	for	7.46	hr
Eq_03			
Eq_04			
Eq_05	for	2.93	hr
Eq_06	for	999	hr

Table 30: Initialization of equipment downs

The problem at simulation start is to estimate the future point in time when the equipment will be in an up-state again. Multiple options are available to solve this problem. The first option is to assign a global average downtime for all equipment. The disadvantage of this is that all equipment reaches the up-state at the same point in time. The second solution is to use a stochastic distribution, which requires confidence runs. So the model execution time increases. This is not acceptable, because the short term simulation needs to provide the results within a limited computation time. Another solution is to use historical data for the downtime duration. The average equipment down duration is used to define the expected downtime at simulation start. This simple solution uses the MTTR divided by two. The underlying assumption is that half of the downtime has already passed. This parameter captures the average downtime behavior from history. Another advantage is that the MTTR value is already available for all equipment in the model.

The results of this approach are depicted in Figure 55. The diagram contains the downtime percentage of the online simulation full fab model. The values of the average fab downs are hidden due to non-disclosure agreement. It becomes clear, that the initial downtime problem has been reduced significantly. At simulation start, a small deviation of 28% still exists only for a few hours. This small deviation is acceptable for online simulation. The reason is that a high fluctuation of the average fab down percent per hour exists in the fab.

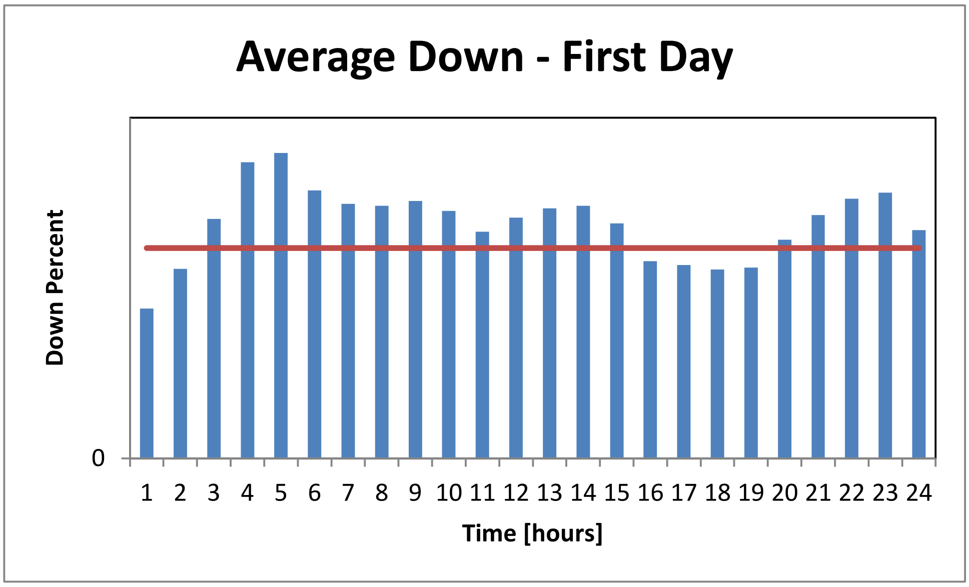


Figure 55: Average down for the first day for the fab model

5.4.7 Remaining Process Time

To initialize the WIP lots properly, another variable is the remaining process time. To compute the remaining process time, the step start time of the lot, the simulation start time, and the predefined process time is used. The remaining process time is the process time of the lot minus the time duration between the process start of the lot and the simulation start.

$$t_{ProcRem} = t_{Proc} - (t_{SimStart} - t_{StepStart})$$

For the beginning of the simulation run, an unusual effect becomes obvious. The generated arrival forecast value is too high (Figure 56). To figure out the reasons for that, it is necessary to have a closer look. The reason is that too many lots move in the first second of simulation to the next operation. Figure 56 depicts the lot moves in the first, second, and third minute. The moves are also available for the first, the second, and the third seconds. It can be seen that during the first minute over 300 lot moves occur. In reality the number of moves is much lower. For the second and third minute, the simulation behavior reflects reality. To have a closer look at the first minute of simulation, the first seconds after simulation start are depicted as well. It becomes clear that from the second and the third seconds onwards, there are almost no lot moves in simulation and in reality. Only for the first second of simulation the number of moves is high.

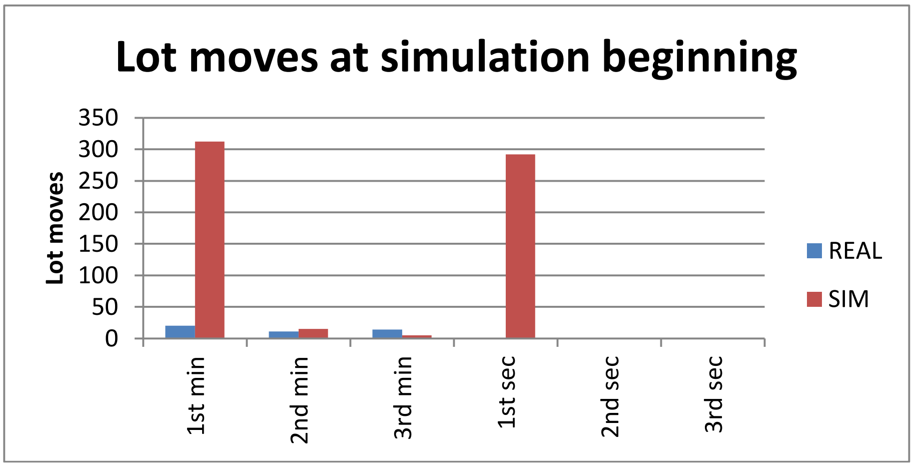


Figure 56: Lot moves during the first seconds and first minutes of simulation

To figure out the reasons, it is necessary to compare the lot trace information that links simulation and reality. This analysis reveals several reasons. Therefore it is necessary to classify the reasons, according to Figure 57. In total there are 292 lot moves during the first second of simulation. 249 out of those lot moves are caused by a very long process time in reality. This means in reality the current process time of the current lot is much longer than the average value for the process time. So in simulation where the average process time is used, the lot should have finish processing already. The result is that all of these lots finish processing at the first second. Many reasons exist for the process time variation. Examples are equipment downs, chamber downs, and slowdowns from overlapping processes at cluster tools.

Besides the process time variance, other reasons exist as well. Real modeling issues only apply for 43 lots. Out of these lots, only 9 lot moves are caused by an unscheduled down of a station. In simulation a lot move, due to an unscheduled down, is counted as a lot move for the equipment. In reality a lot which repeats the operation is counted only once within the equipment. For 3 lots, the small lot size is the reason. The process time is much longer in reality than it is in simulation (Morrison 2011).

Another reason is the data quality. Therefore 2 lots finish too early because the process time is not available for the particular combination of product, process and equipment. A generic process time average is used for the equipment. This average process time is too small compared to the product and the process specific process time. So the lot finishes too early. The remaining 29 initial lot moves are also caused by bad data quality. In these cases, the average process time in simulation does not reflect the average process time in the fab. The process time values from the fab data source are not correct.

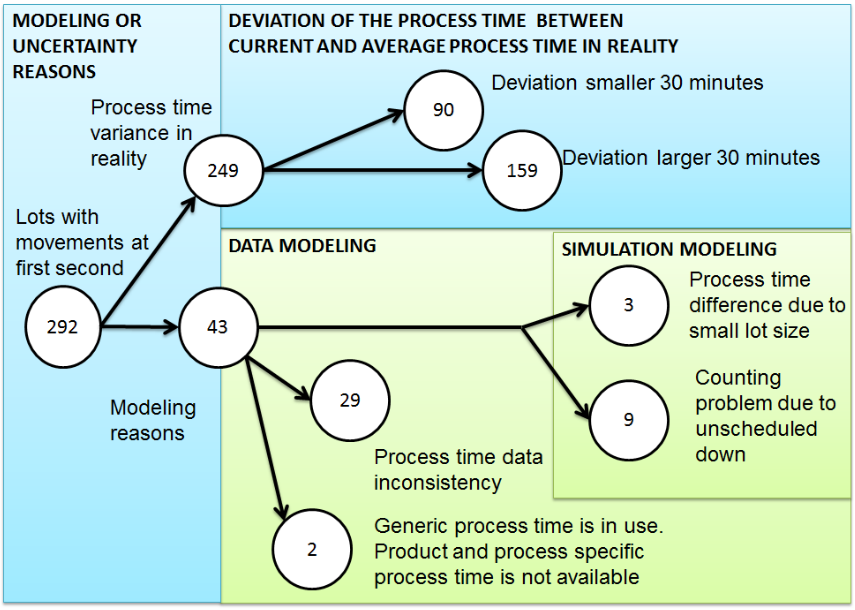

MODELING OR UNCERTAINTY REASONS
DEVIATION OF THE PROCESS TIME BETWEEN CURRENT AND AVERAGE PROCESS TIME IN REALITY
Process time variance in reality
249
90
Deviation smaller 30 minutes
159
Deviation larger 30 minutes
Lots with movements at first second
292
43
Modeling reasons
DATA MODELING
SIMULATION MODELING
3
Process time difference due to small lot size
9
Counting problem due to unscheduled down
29
Process time data inconsistency
2
Generic process time is in use. Product and process specific process time is not available

Figure 57: Classification of initial lot moves

The conclusion is that most of the lot moves are caused by process time variance in reality. Several lot moves are also caused by bad data quality. Only a few lot moves are caused by simulation modeling reasons. The impact of deviations, caused by the simulation model itself is quite low, because only few lots are affected. It is highly useful to manage the process time variance and to improve the data quality.

Multiple solutions are applicable to avoid lot moves at the first second in simulation. One solution is to delay the lots with a predefined percentage of the process time. For example all lots, which will finish at the first second of simulation, remain in the process for another 5% of their process time. This simple solution smoothes the problem but it does not resolve it. In fact at simulation start, the exact point in time when the process will finish processing is not known. Therefore the gap between simulation and reality still exists.
Another solution is applicable for those lots that finish early (Figure 58). It is clear that each process has a particular process time distribution, where in simulation the average process time is used. If a lot exceeds the average process time, this lot finishes processing at the first second of simulation. To avoid this release, it is also possible to apply the remaining part of the process time distribution for those lots, see Figure 58. The red area describes the remaining process time distribution. The lots follow their remaining process time distribution accordingly.

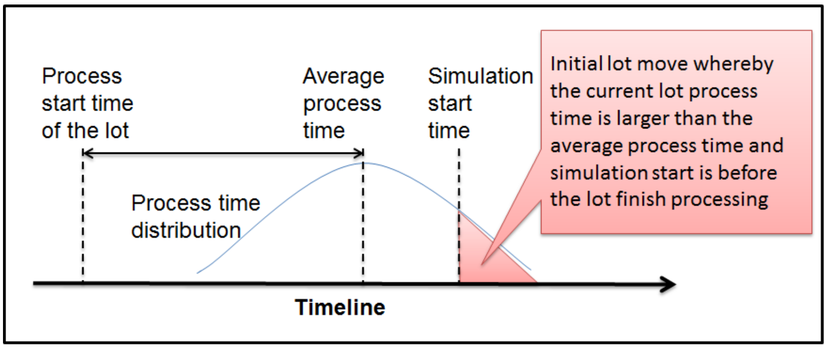

Process start time of the lot
Average process time
Simulation start time
Initial lot move whereby the current lot process time is larger than the average process time and simulation start is before the lot finish processing
Process time distribution
Timeline

Figure 58: Process time distribution and simulation start

The disadvantage of this solution is that confidence runs are required. The point in time where the lot finishes processing highly depends on the stochastic distribution. Another disadvantage is the high effort undertaken for data collection and computation time. At the moment a single integer process time value is used for each combination of product, equipment, and process. More than 100000 integer values exist. For the new solution it is necessary to provide a distribution for each value. This solution is not feasible and out of scope for online simulation.

From application perspective the effect of the moves at the first second is not highly relevant. The stated effort to implement a solution is also very high. This is also a reason why none of the solutions have been applied. The negative effect is small because the level of detail for the reporting of the lot moves is measured in days and not in seconds. So the short warm-up time is not transparent. Most of the lots that finish at the first second of simulation finish early in reality, within the first day. So the statistics for simulation matches reality with a fairly small deviation.

From scientific perspective the effect of the first second is most interesting because the first second in simulation is the transition between a correct initialized static model and uncertain dynamic simulation run. Most publications refer to the steady state whereby the effects during the warm-up period are not widely researched.

5.4.8 Determine the Warm-Up Period

The analysis above shows that the warm-up period affects the simulation results for online simulation. The impact on the simulation results depends on the particular modeling feature.

In literature the standard way so solve this problem is to cut off the warm-up period (Robinson 2002). It is used for steady state simulation models, without a proper initialization. One way is to use the fab WIP to determine the warm-up period (Mahajan and Ingalls 2004). The following example shows the effects of a good and bad initialization on the fab WIP. In Figure 59, the fab WIP is depicted for two simulation scenarios. The assignment of WIP lots and the current lot operation is part of the model initialization for both scenarios. In addition, the “SIM_INIT” scenario contains the initialization of the lot hold, the lot rework, the current equipment of the lots, and the initial equipment states. The “SIM_NO_INIT” scenario does not contain these additional features. Due to non-disclosure agreement, the fab level is depicted in percent, whereby the reference point of 100% is the value of the first days from the first scenario.

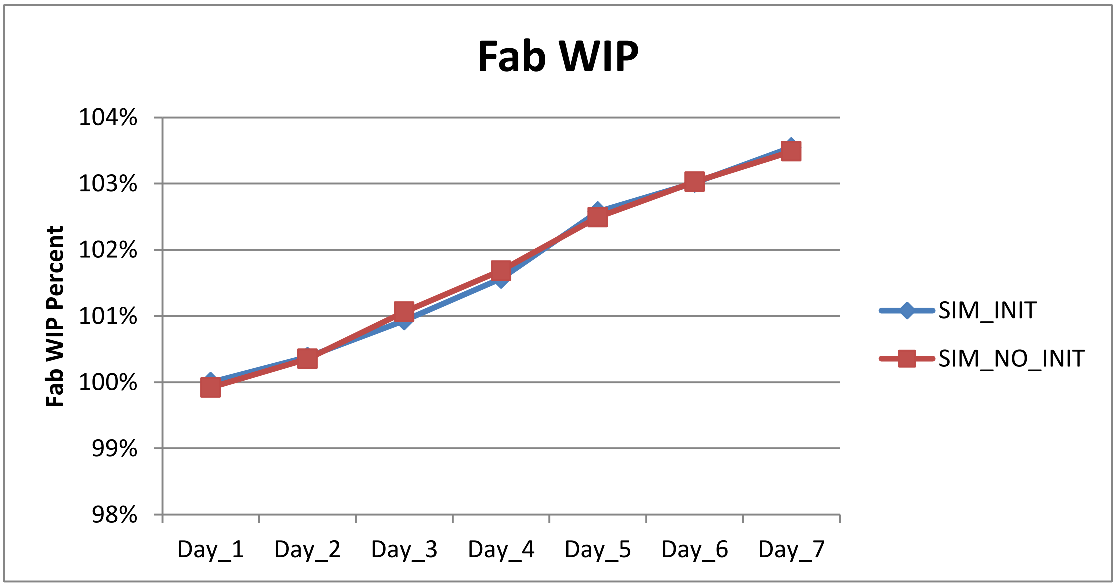


Figure 59: Comparison of fab WIP for an initialized and non-initialized Scenario

As seen in Figure 59, the fab WIP is almost identical, even if the model initialization is very different. The effects of a detailed model initialization have no significant effect on the fab WIP. It is clear that this way to analyze the fab WIP cannot be applied to determine the warm-up period.

Due to the transient nature of the fab it is hardly possible to determine a warm-up period for a short term fab model. Even if the length of the warm-up period is known, it is not useful to cut off the warm-up period. It contradicts the concept of short term simulation. Due to this fact, the determination of the warm-up period does not have a lot of value. Therefore it is not part of this thesis. The focus is to analyze the individual model behavior at simulation start. These analyses contribute to the improvement of the model behavior at simulation start.

5.4.9 Conclusions for Model Initialization

The conclusion is that even very well initialized simulation models have a warm-up period. During this period the model behavior is different from the behavior of the real fab. The model behavior during this warm-up period is also different from the model behavior after the warm-up period. For short term simulation it is not feasible to cut off the warm-up period. So for this research the reasons for the deviation of the warm-up period have been analyzed. From a practical perspective, the objective is to minimize the impact of the warm-up period. As long as the effort supports the accuracy benefit of the simulation, it is useful to enhance the model initialization.

Different analyses identify the model elements, which affect the warm-up period:

- Incorrect sampling which will be applied twice
- Initial lots enter next operation
- Initial hold lots leave the hold state
- Incorrect batch initialization
- Estimation of initial equipment downtime duration
- Estimation of first equipment down

Form research perspective, the conclusion is that in the future additional effort is required to analyze the warm-up period. The objective is to further reduce the warm-up effects, especially for models with a very short time horizon of less than one day. With an even more detailed initialization, the expectation is that the model accuracy increases even further. For steady state simulation models, a lot of literature is already available to determine and to cut off the warm-up period (Hoad et al. 2008). It is useful to provide multiple solutions in literature with the objective to reduce the warm-up period as well.

5.5 Model Validation & Verification

The focus of this section is the model validation process for online simulation. The model validation is important because it guarantees a good quality level regarding the forecast accuracy and the usefulness of the model. Beside the data integration the validation part takes up the most effort within the online simulation project. This section presents several model validation methods from literature. According to these methods several examples from this project are available in this section. Furthermore the modeling gaps between simulation and reality are presented. The way to overcome these gaps is also shown. The concurrent model validation and implementation process is described. The description includes the specific needs of a high detailed model initialization and the effect of the whole validation process. The work center validation process is described in detail.

5.5.1 Reporting for Model Validation

The objective of model validation reporting is to reveal the weakness of the simulation forecast to improve the simulation results. Therefore the reports need to show the deviation from reality. It is useful to find powerful KPI with an appropriate level of detail to show that difference. The purpose is to narrow down the reasons of modeling gaps and to find improvements. Therefore the reporting for model validation is very different compared to the reporting for the user of online simulation. For the validation of online simulation, several examples of reporting elements are presented:

- Fab and work center performance
- Model initialization
- User reporting parameter
- Execution times
- Error log

Multiple chart types are used to capture the model accuracy. Examples are pareto-charts, histograms, simple line charts, cumulated charts, and many other chart types.

Fab and Work center Performance

Several custom reports are in use to validate the fab model and the underlying work center models. These reports depend on the particular purpose. For the model validation phase, three examples are used to demonstrate the interaction of the reporting and the model validation:

- Cycle time of a lot
- Process time of a lot
- Work center WIP

During the model validation phase, the cumulated cycle time chart is used to show cycle time differences for a particular route (Figure 60). The chart shows operations with a cycle time deviation. It also reflects the general behavior of the simulation model. During model validation it turns out that the simulation model is too optimistic.

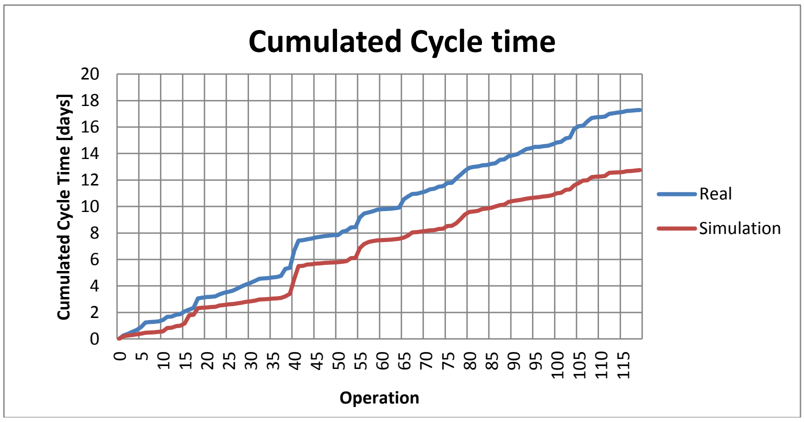


Figure 60: Cumulated cycle time comparison for one route

To validate the process time, an option is to compare the process time of a single lot. Figure 61 displays the raw process time (RPT) for simulation and reality for 160 executed operations. The largest RPT for simulation and reality is on position 0. The smallest RPT is on pos. 160.

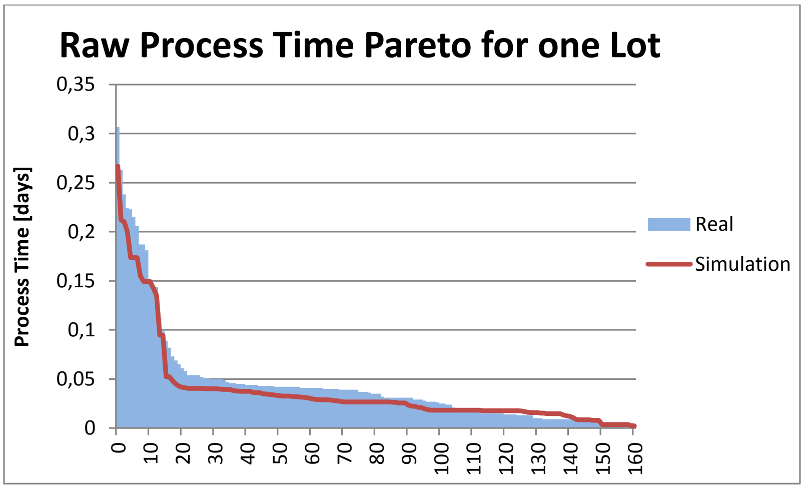


Figure 61: Pareto RPT for one lot

To bring model validation down to work center level, it is useful to compare work center performance in terms of average work center WIP. Figure 62 depicts the work center WIP difference between simulation and reality. With this chart it is well illustrated which work center WIP is too low or too high.

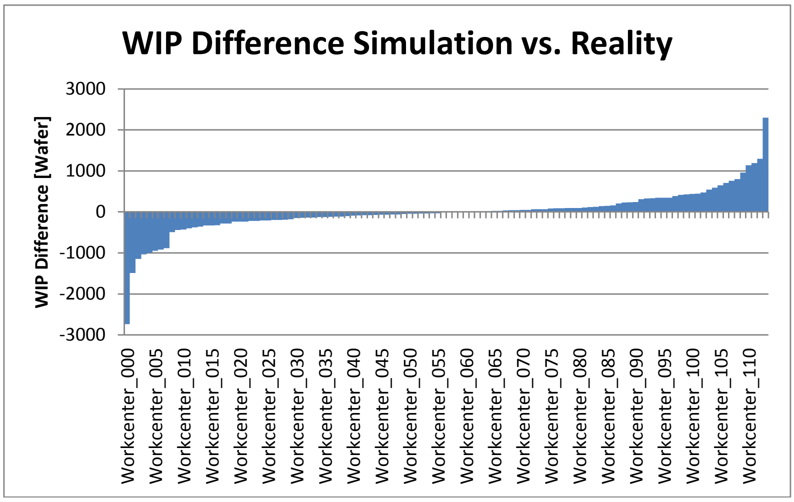


Figure 62: Pareto of work center WIP difference between simulation and reality

Another idea is to use the lot based matching approach. It compares the cycle time of the lots. The data source is the lot trace for simulation and reality. The idea behind is to compare only a few parameters to figure out whether the results of the whole simulation model are improving. It is not necessary anymore to compare numerous parameters for several work centers to see if an individual model feature improves the overall forecast quality of the whole fab model. Just two parameters are relevant:

- Matching ratio
- Deviation of the cycle time

The matching ratio is the percentage that indicates how many lot movements from the simulation model match with the lot moves in reality. The cycle time deviation parameter compares the cycle time for all lots. It shows the average time difference for the lot arrival events, between simulation and reality.

The advantage of the cycle time deviation parameter is that only a few parameters are necessary to determine the behavior of the whole fab. Different categories can be derived, e.g. the cycle time deviation per day, per work center, per product. The disadvantage is that not all lots find a corresponding matching partner. Reasons are for example sampling, rework, or incomplete lot release information. If a different lot is chosen in simulation at a sampling operation, the matching ratio is lower, even if the work center performance is not affected.
To overcome this disadvantage, the matching ratio parameter is necessary.

Model Initialization Reports
To validate if the model initialization is correct, several charts are required to compare the model with the real fab state at simulation start. One example is available in Figure 63. The chart compares the initial WIP wafer in simulation and reality. The figure compares only those wafers from productive lots.

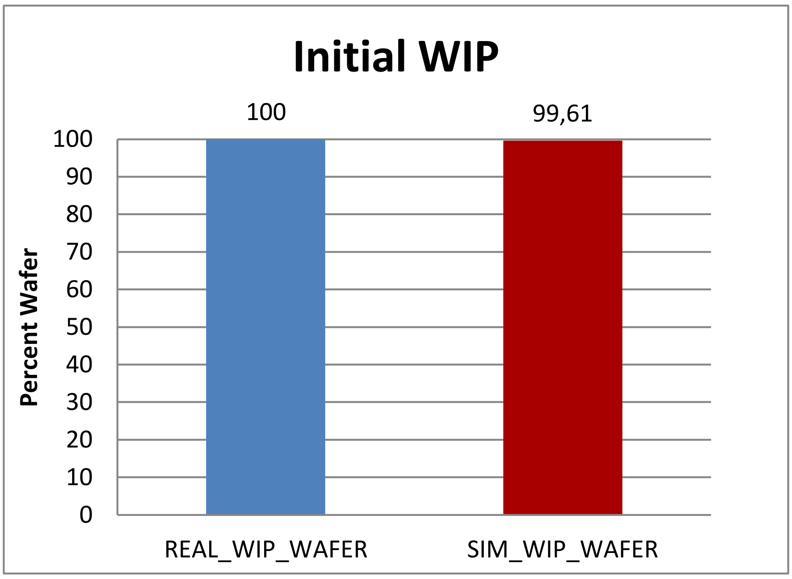


Figure 63: Initial WIP wafer

As seen in the previous sections, many other reporting charts have been used to evaluate the model quality for model initialization. Therefore this section contains only one example.

User Reporting Parameter

To increase the forecast accuracy, it is necessary to consider the user parameter for model validation as well. The purpose is to monitor the forecast results from user perspective. Therefore the system generates the reports for cycle time, WIP, wafer arrivals, and wafer moves. Figure 64 to Figure 67 contain the results not only for simulation, but also for reality. After seven days model validation is applicable when the results from reality are available.

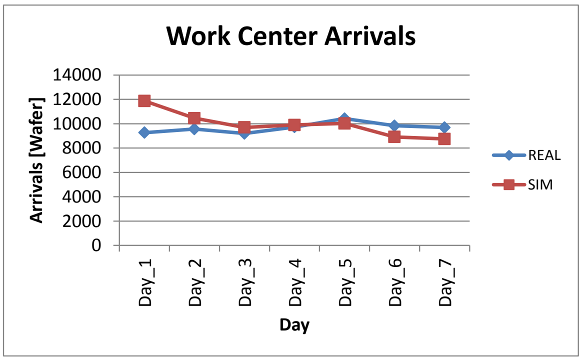


Figure 64:Work center validation arrivals

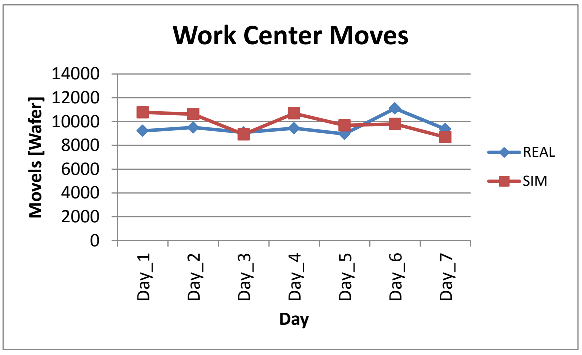


Figure 65: Work center validation moves

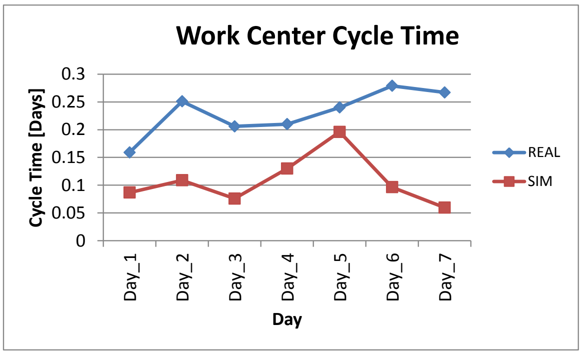


Figure 66: Work center validation cycle time

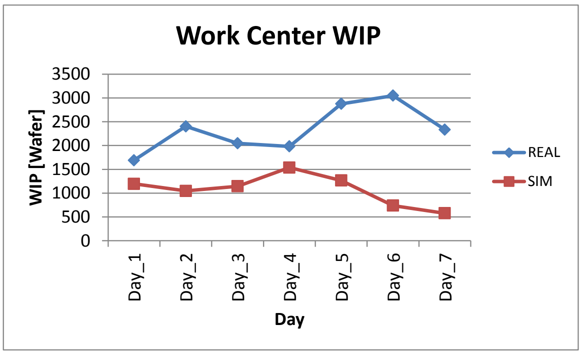


Figure 67: Work center validation WIP

Execution Times

One requirement of online simulation is the real time capability. Therefore the runtime of the forecast and its components are important (Table 31). During model validation the computation time of each component is in the range of interest. In different phases of the project these execution times vary significantly. Due to new features in the model and computation time reduction strategies in the source code, the computation time is fluctuating.

Component	Execution time
Fab data extraction in database	2 min fab export (without master data)
Time for cleaning and error handling	2 min
Data transformation into target schema	1 min
Export of simulation files on hard disc	1 min
Import of simulation files into memory and model initialization	5 min
Simulation runtime	5 min

Table 31: Computation times

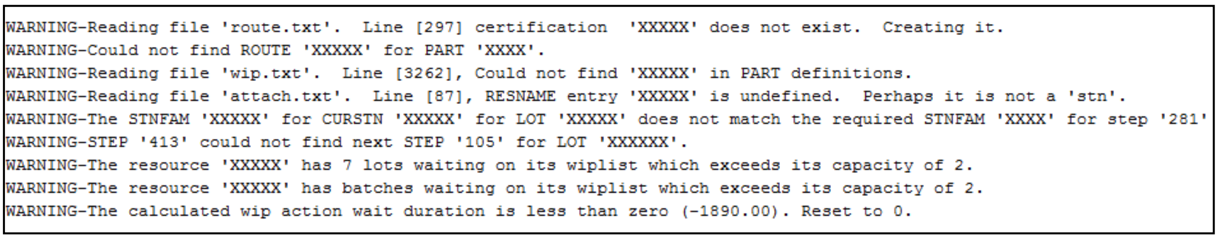
The objective is to reduce the computation time for providing forecast results. These time values are used to identify which components take up major proportions of the total time needed. The components with the highest duration provide the best opportunities to reduce the total time.

Error Log

Log files are an important factor to identify and to fix errors. Within the whole online simulation system two main logging systems exist:

- Database error logging
- Simulation model error logging

The error log of the database is a result of the data analyzer module. The error log is highly recommended to reveal wrong data input from the data sources of the fab. It becomes necessary to correct these data errors. The simulation error logging is useful to find system bugs and input data errors that have not yet been fixed in the database. An example of such a logging file, created during the model validation phase, is available in Figure 68.



```
WARNING-Reading file 'route.txt'.  Line [297] certification  'XXXXX' does not exist.  Creating it.
WARNING-Could not find ROUTE 'XXXXX' for PART 'XXXX'.
WARNING-Reading file 'wip.txt'.  Line [3262], Could not find 'XXXXX' in PART definitions.
WARNING-Reading file 'attach.txt'.  Line [87], RESNAME entry 'XXXXX' is undefined.  Perhaps it is not a 'stn'.
WARNING-The STNFAM 'XXXXX' for CURSTN 'XXXXX' for LOT 'XXXXX' does not match the required STNFAM 'XXXX' for step '281'
WARNING-STEP '413' could not find next STEP '105' for LOT 'XXXXXX'.
WARNING-The resource 'XXXXX' has 7 lots waiting on its wiplist which exceeds its capacity of 2.
WARNING-The resource 'XXXXX' has batches waiting on its wiplist which exceeds its capacity of 2.
WARNING-The calculated wip action wait duration is less than zero (-1890.00). Reset to 0.
```

Figure 68: ASAP error log, anonymous

5.5.2 Methods of Model Validation & Verification

The range of interest for model verification is to check whether the implementation of the online simulation model complies with its specification (Balci 1986). The range of interest for model validation is to check if the model is functional. The model is useful for online simulation, if the results represent the system with sufficient accuracy. For this project the model validation is highly important. The first reason is that it is not practical to specify the whole fab with all of its details at the beginning of the project. The system of the wafer fab is very large and very complex. During the execution of a long term project a lot new knowledge will be acquired. The online simulation project is also a scientific project. The quality of the results and the elements, which have impact on these results, are not known beforehand. To ensure a high model quality, many validation and verification methods from literature have been used (Rabe et al. 2007, Balci 1998, Sargent 2005):

.

- Compare simulation results with reality
- Sensitivity analysis
- Interview with production departments
- Boundary value testing
- Review
- Monitoring
- Statistical methods
- Module testing
- Trace analysis
- Cause effect analysis
- Debugging

In this project, for example the trace analysis is used. It checks, whether the split and the merge of the send-ahead feature is effective in simulation. Module testing has been applied. To test new model features like Kanban, small models are used at the beginning. The individual work center model behavior is tested with small models as well and the real lot arrival stream. Parts of the model have been crosschecked with other (simulation) models which are in use at the customer's side. A major part of the V&V effort is the comparison of simulation results to reality. For this comparison the level of detail changed from fab level down to equipment level. The input data also changed. For a real forecast scenario no future data is used. For a validation scenario future data for unscheduled equipment downs is used. Therefore it is possible to exclude vast amounts of uncertainty and focus on systematic problems.

5.5.3 Model Validation for Manual and Automated Model Generation

To realize online simulation, first the model has been created manually. This model is useful as a template to develop the automated model generation (Figure 69).

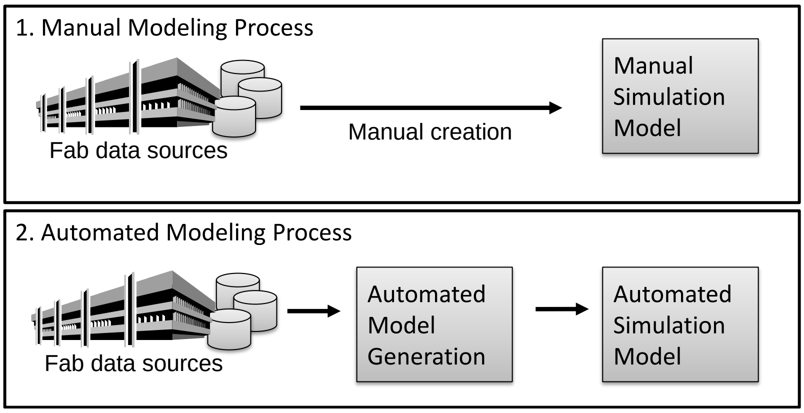

1. Manual Modeling Process
Fab data sources
Manual creation
Manual Simulation Model
2. Automated Modeling Process
Fab data sources
Automated Model Generation
Automated Simulation Model

Figure 69: Sequential manual and automated modeling process

The manual model is useful to figure out if online simulation is feasible. It minimizes the risk of the project failing because this information is available at a very early stage. Another advantage is being able to estimate the effort and to see which results accuracy can be expected. This manual model uses only fab data sources to see whether it is possible to create a fab model automatically with little manual interaction. This model considers the requirements of online simulation for short term prediction. The validation methods of the manual model are equivalent to common fab model validation methods. This will not be elaborated any further.

The second part contains the automated model generation based on real fab data. The basis of the automated model generation is the specification of the manual model. The required fab data sources are already identified. The target simulation model format is also clear beforehand. From validation and verification perspective, the manual model has significant advantages. It is used as a reference model for the automated model. A direct comparison of model elements displays many differences.

5.5.4 Work center Model Validation

The model validation on work center level is very important because the main results of online simulation are results on work center level. Examples are work center WIP and work center lot arrivals. So a comparison of work center results from simulation and reality improve the model quality. Several KPI, especially lot arrivals and lot moves are in use. Figure 70 indicates important validation decisions based on work center lot arrivals and lot moves. The overall target is to reach good predictions for lot moves and lot arrivals (E). If both parameters are similar in simulation and in reality, then the work center modeling is correct. If the work center arrivals deviate, then the origin of the problem is either the sampling rate or the throughput of a preceding work center (A and B). If the lot arrival is correct but the moves do not match reality, then the work center throughput is not correct (C and D).

Arrivals \ Moves	Moves High	Moves OK	Moves Low
Arrival High	A	A	A and D
Arrival OK	C	E	D
Arrival Low	N/A	B	B

Legend: 🟦 Upstream Workcenter
🟧 Current Workcenter

Figure 70: Decision matrix for work center validation

A) Upstream work center moves are too high or current work center sampling rate is too high
B) Upstream work center moves are too low or current work center sampling rate is too low
C) Current work center throughput is too high
D) Current work center throughput is too low
E) Current work center behavior reflects reality

Based on this matrix, further actions are required. If the arrival is not correct (A and B), the first element which needs to be checked is the sampling rate of the current work center. If the sampling rate is correct, the origin of the problem is a preceding work center. It makes sense to select the preceding work center, which causes the problem. The idea behind this approach is to improve the current work center behavior, by improving its arrival rate. In a case where the current work center arrival is correct, but the moves do not reflect reality (C and D), it is useful to analyze the current work center in detail.

The detailed work center analysis contains the comparison of several work center properties in simulation and reality (Figure 71). First it is useful to identify which equipment is related to a work center. In some cases two work centers share equipment or the equipment is not in productive use any more. The second test compares the equipment states in simulation and reality in front of different time horizons. The objective is to figure out if the equipment state matches reality. In many cases, an unscheduled down event causes the deviation between simulation and reality. With regard to the dedication it is useful to compare the equipment moves in simulation and reality for each process. Differences in these lot moves indicate that the equipment dedication is not correct. A major reason is the complexity of the dedication mechanisms. In several cases the dedication data is not up to date. Another test case is the comparison of the basic data, like throughput and process time. In many cases a deviation exists due to data quality problems. Custom work center properties like setup rates, batch sizes, and custom dispatch rules need to be checked as well. All these factors help to narrow down the reason of the forecast discrepancies for a particular work center.

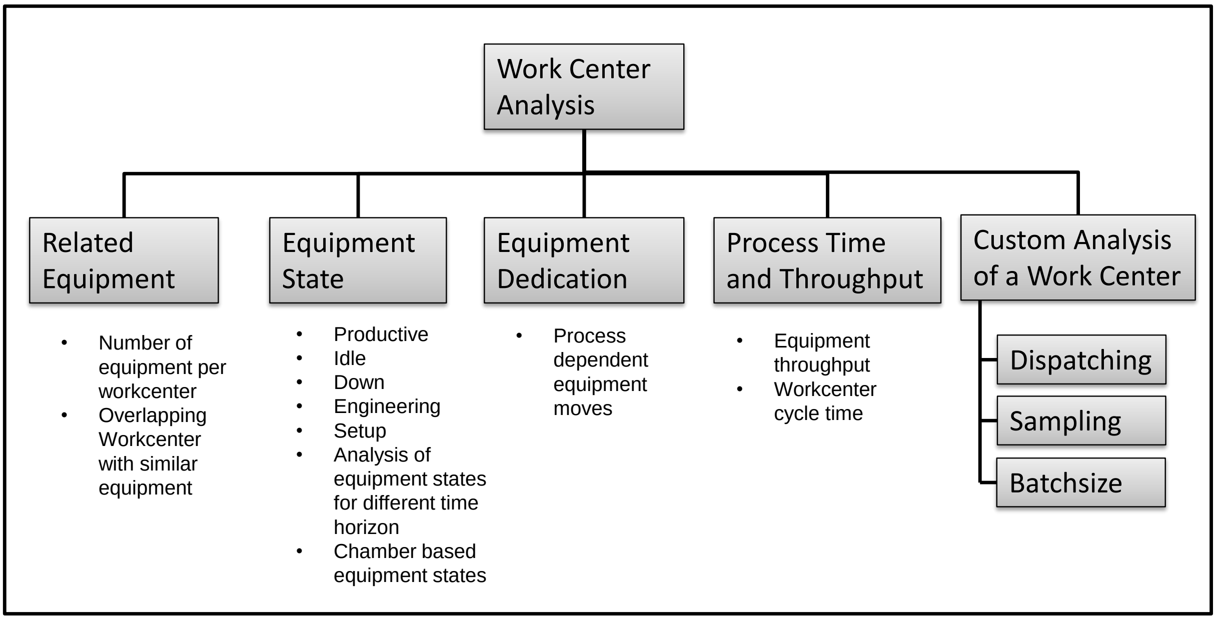

Work Center Analysis
Related Equipment
Equipment State
Equipment Dedication
Process Time and Throughput
Custom Analysis of a Work Center
Number of equipment per workcenter
Overlapping Workcenter with similar equipment
Productive
Idle
Down
Engineering
Setup
Analysis of equipment states for different time horizon
Chamber based equipment states
Process dependent equipment moves
Equipment throughput
Workcenter cycle time
Dispatching
Sampling
Batchsize

Figure 71: Comparison of work center elements in simulation and reality

Figure 72 to 75 depict the results of a single work center during the model validation process. In Figure 72 the WIP in simulation is growing, while in reality it is quite stable. Obviously the work center throughput is too low. The reason for this behavior is that the equipment capacity is too low. The number of lots that can be simultaneously processed is not correct. The throughput related time value is also not correct. So the equipment throughput was much too low. These problems are caused by incorrect input data. After fixing these two problems, the work center WIP behavior is better, as depicted in Figure 73. It can be seen that the scaling of the diagram changes to portrait more details compared to the previous diagram. In this phase it was observed that the initial WIP of the work center (Day_1) is not correct. The WIP at model initialization has to match reality. Further investigation shows that the work center assignment is not correct. After applying the correct work center assignment, the initial WIP matches reality, see Figure 75.

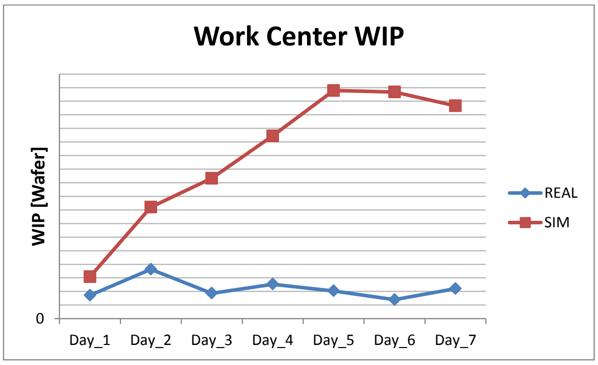


Figure 72: Throughput too low

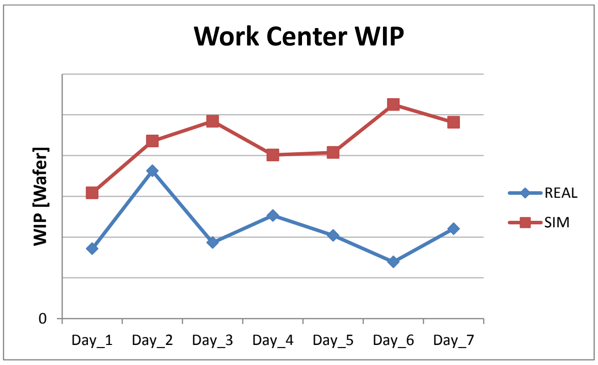


Figure 73: Initial WIP at Day 1 is not correct

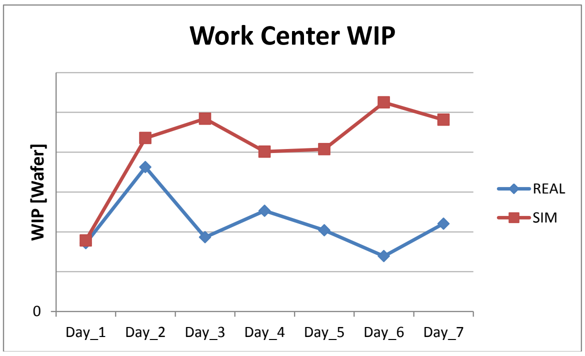


Figure 74: Arrival too high

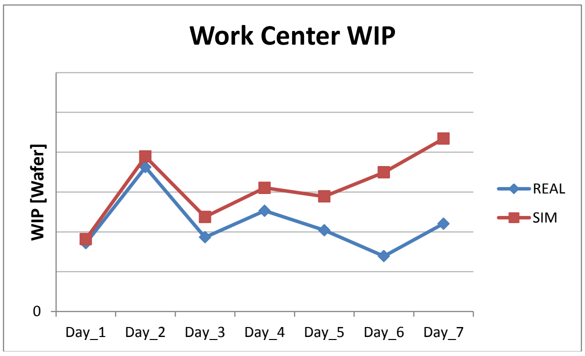


Figure 75: Work center trend match reality

In Figure 74 the work center WIP is still too high because the work center arrivals from an upstream work center are too high. With such results, the target is to analyze and to fix the performance of other work centers in the fab. After several other work centers have been analyzed, the current work center behavior improves significantly due to an improved arrival pattern. In Figure 75 the work center WIP trend matches reality, especially during the first days. At the end of the simulation time horizon the particular work center WIP deviates from reality. For this particular case, the forecast is sufficient to predict the work center behavior for the next 4 days. To extend the simulation horizon to five days or more, further model validation and model improvements are required.

This example demonstrates that various error reasons reduce the model accuracy. During simulation model validation about 25% of the faults are caused by modeling issues, where the concept or the implementation the simulation model is not accurate. An improvement of the concept and the implementation solves these problems. About 25% of the faults are caused by a wrong interpretation of the input data. Those cases have been corrected in the fab data interface. Another 50% of the errors are caused by incorrect and inconsistent input data. To solve these problems, the automated error handling module has been modified to correct data value for future models. In several cases the underlying data values in the wafer fab has been updated.

5.5.5 Basic Validation Principles for Online Simulation

Online simulation has custom requirements for the model validation process. The two major directions for model validation are the following:

- Top down approach (Level of detail)
- Initialization to full period (Simulation horizon)

For the online simulation project, Figure 76 shows that the model validation process starts at an abstract level with the objective to increase the level of detail. This approach is quite common. For the full fab model it is not useful to start the model validation with the highest level of detail if major modeling issues on fab level have not been fixed yet. For the full model, the validation process starts on the fab level, continues with the department level and works its way down to the work center level.

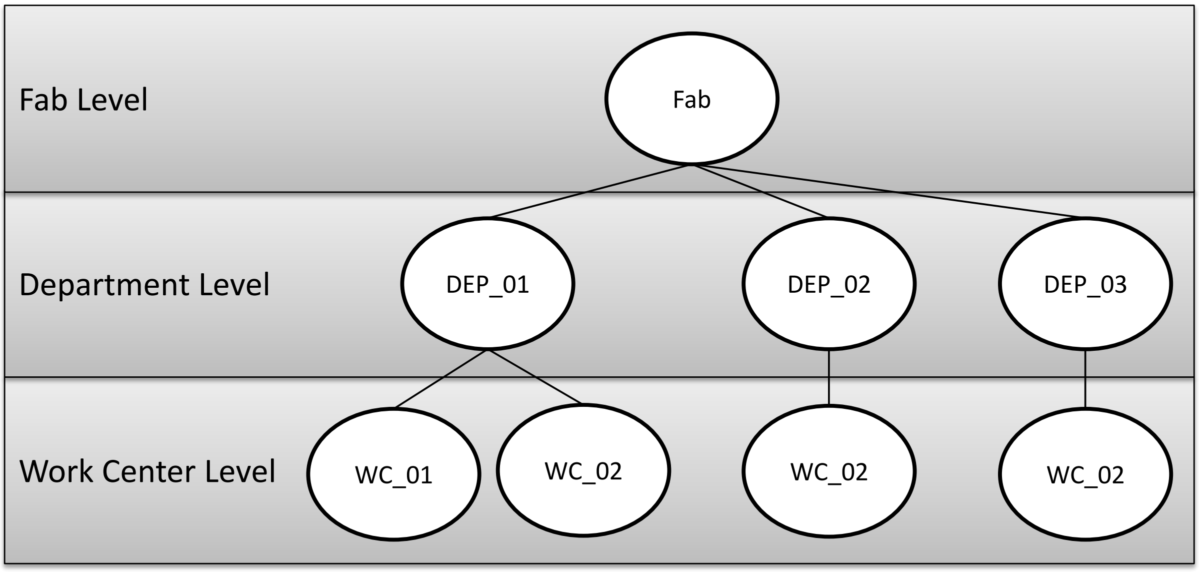


Figure 76: Top down model validation approach

To meet the needs of short term simulation it is very useful to supplement this approach with the major requirement of a detailed model initialization. The expectation is that the forecast results highly depend on the initial state. It is necessary to consider this fact for model validation. Only if the initial state is correct, it is useful to extend the time horizon for model validation. The validation process analyzes the initial state, the first second of simulation, the first hour, the first day, and finally the whole simulation forecast horizon. Figure 77 extends the top down model validation approach to the time horizon. Starting point is the fab level and the initial state. The target is to have a detailed work center forecast for the full time horizon.

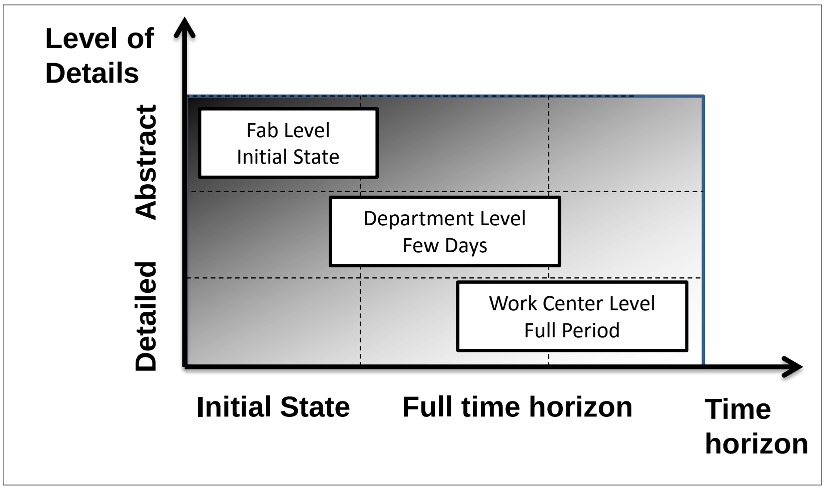


Figure 77: Validation process: Level of detail & Simulation Time horizon

Examples for this process are available below. Starting point is the fab level at the initial state Figure 78. The target of model validation is to achieve good forecast results on the work center level for the full simulation time horizon.

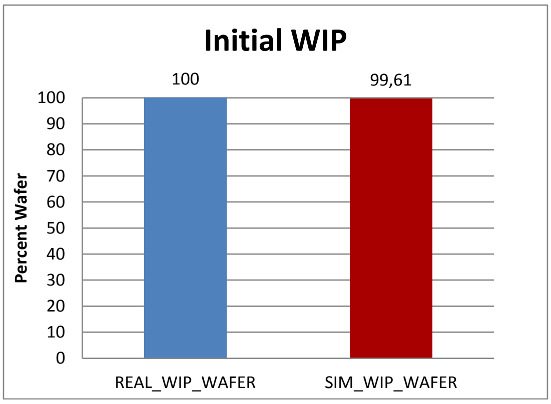


Figure 78: Initial fab WIP

On fab level at the initial state it is one important precondition to check whether the current WIP matches reality. It is useful to identify modeling gaps. The effect of inconsistencies for the full time horizon and for the detailed work center analysis is enormous. Possible reasons for an initial fab WIP deviation are the following:

- Data inconsistencies: Lots in simulation do not have valid routes or part definitions
- Non-productive lots: Development, test and engineering lots, affect the statistics collection

Different KPIs, like cycle time, lot release and wafer out are also used for fab level model validation. It is effective to validate the full time horizon on fab level to see how good the average performance is.

According to the top down approach, the department level follows the fab level. A model validation example is depicted in Figure 79 and 80. The WIP initialization matches reality. After 1 day of simulation, a WIP deviation exists for department DEP_06 and DEP_09.

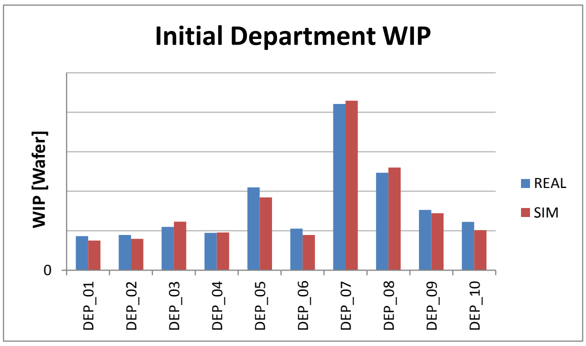


Figure 79: Department WIP at simulation start

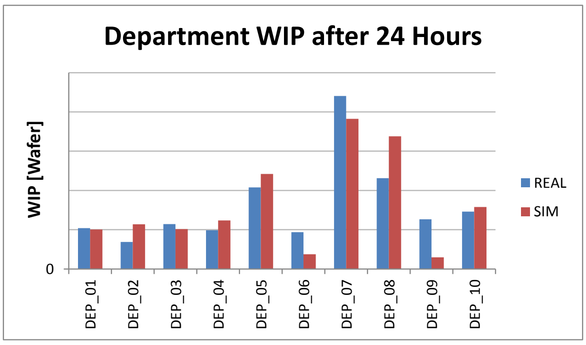


Figure 80: Department WIP after 1 day

For departments where the WIP deviates, it is necessary to have a closer look at the related work center. One example is depicted in Figure 81. The reason for the deviation of the department is the deviation of a single work center.

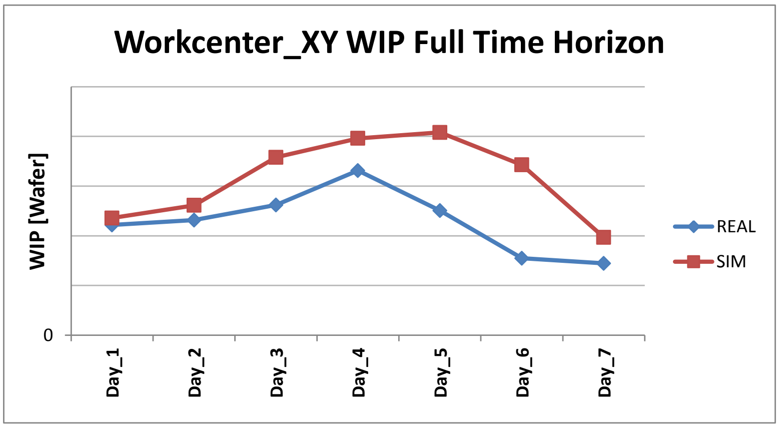


Figure 81: WIP for a single work center for the full simulation horizon

The model validation part identifies many reasons for work center deviation. The common reasons for these gaps are throughput mismatch, abstraction of dispatch rules, hold, rework, and sampling.

5.5.6 Concurrent Model Validation and Development Process

The model validation process goes hand in hand with the implementation of online simulation. One objective of model validation is to identify errors and to constitute modeling improvements. These improvements need to be implemented for the automated model generation. The question is how to synchronize the validation part and the implementation part. The whole process is highly critical, especially when a team is involved to execute the implementation task and the model validation task with different members of staff. The challenge is to progress fast and to avoid dependencies or idle times for the whole team. For this scientific online simulation project the uncertainty is high and complex fab relations exist. The model validation process of the automatically generated model starts when the automatically generated model is available. All required features, based on the manual model, are already implemented.

Figure 82 depicts three different validation and development processes. The first validation and development process in Figure 82 is a sequential process (1). The timeline is available for the current auto generated model and the validation model. The first action is to copy the current automated model and use it as a validation model (A). The second step is to identify the problems in the validation model and find appropriate solutions. For testing it is necessary to apply changes for the validation model manually. The third step is to communicate the desired solution and implement the required changes in the automated model generation module (B). After the implementation and the testing of those changes, the next auto generated model is up to date. This new model is used for further validation. The disadvantage is that during the time of the model validation the implementation process is idle. During the implementation phase, the model validation does not proceed.

The parallel model validation process improves this situation (2). The idea is to get an auto generated model once (A). Afterwards a single model is used to identify multiple modeling problems. For the validation model the changes for (B, C, D, and F) are applied manually.

If a problem has been fixed for the validation model, it is necessary to implement the same changes (B, C, D, and F) for the automated model generation. So the model validation continues, even if a problem has not yet been fixed for the automated model. One disadvantage is the high effort to apply changes in the manual model and in the automated model. A big problem with this approach is that the validation model and the automated model deviate after a while, especially if many elements have been amended. The reason is that many changes are not 100% identical in both models. The models are not synchronized. The manual model is outdated as well because it refers to a time period in the past.

The solution is the synchronized model validation approach (3). Some issues are identified with a single manual model (B, D). The issues are implemented parallel in the automated model generation module. For the next auto generated model (C), the issue (B) is already implemented while (D) is not. Further model validations proceed (E, F). The changes for (B) are synchronized for both models. The model validation team decides when it is useful to work with the next auto generated model. This decision is based on the quality of the validation model and the auto generated model.

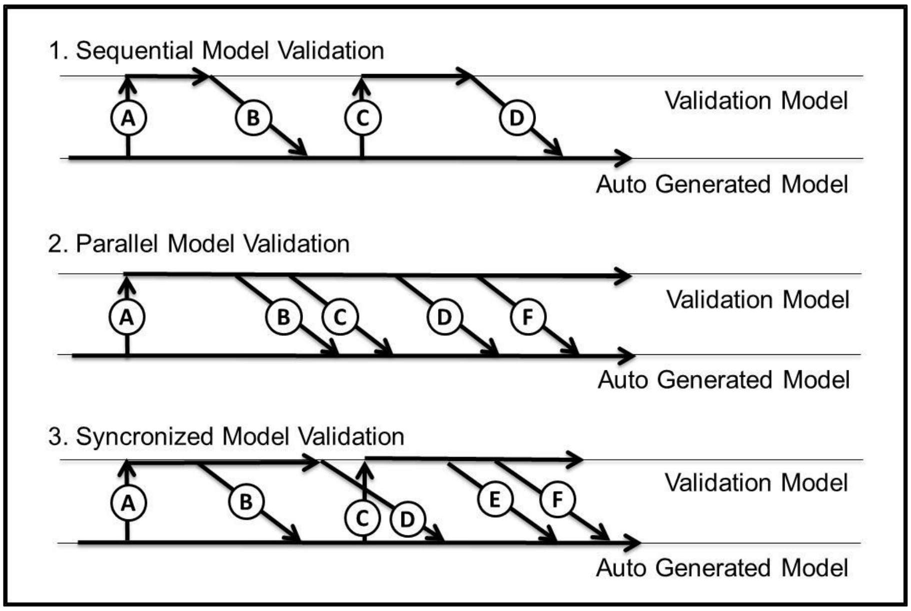


Figure 82: Model validation process for automated model generation

The model validation process is very time consuming. Therefore it is necessary to reduce the time for the single validation loop, which includes the problem identification, the manual application of different solutions, the decision for a solution, the implementation in the automated model generation, and the testing of the results. During the whole project, the time for such an iteration reduces decisively.

At the beginning of the model validation phase, the second approach is used. It is possible to work parallel on the model validation and on the implementation. Many critical issues have been identified with a single model. It takes only little time to identify errors, but it takes a long time to incorporate most of the changes in the automated model.

At the end of the model validation phase, the identification of improvements takes much more time. The third approach is used to synchronize both models often. At the end, the time duration between the identification of an issue and the related implementation for the automated model was considerably shorter.

5.6 Conclusions and Outlook

The online simulation model is capable to provide detailed forecast results very fast. It reaches a high degree of automation. The conclusions and outlook section describes the facts and achievements of the online simulation model. For future research the automated model validation is an attractive research area.

5.6.1 Modeling Facts

The simulation model contains more than 100 work centers from the fab, with more than 700 pieces of equipment. Almost 100 different routes are part of the model with over 400 operations per route. Nearly 200 products exist. The model contains a few thousand lots, including WIP lots and future lot releases. The simulation model execution time is about 5 minutes for a single run forecast scenario. The model contains multiple features. Examples are rework, hold, lot attribute dependent sampling, dedication, setup, equipment down, and send-ahead lots. Custom dispatching rules mimic the real dispatch behavior. A major part of the model is the simulation model initialization. It is initialized with many details for the current lot state and the current equipment state. The model reflects the transient behavior of the wafer fab. The simulation model is running with ASAP simulation software. The model size is about 65 MByte. The simulation forecast horizon is one week.

5.6.2 Conclusions

In order to develop online simulation the first step is to describe the requirements of online simulation.

The key factors of online simulation are a high speed, a high accuracy, and a high level of detail. The concept section describes how to meet these requirements. To achieve a high speed, a high degree of automation is necessary within all functional parts of online simulation. To achieve a high accuracy, the simulation model contains numerous modeling features. This thesis lists and describes several features of the simulation model, including the particular level of detail. The focus of this thesis is the model initialization. Online simulation highly depends on the initial state, whereby it is not feasible to cut off the warm-up period. This thesis describes numerous features to improve the model quality at simulation start. Examples for those features are the initial equipment state, the initial lot hold, the initial lot rework, the current operation of WIP lots, the current station, and the remaining process time of WIP lots. The model validation is a challenging task. An overview is given of the used methods, the reporting elements, and several aspects of the model validation process itself. Altogether it becomes feasible to generate an online simulation model in semiconductor manufacturing.

5.6.3 Outlook: Automated Model Validation

In Fujimoto et al (2002) H. Szczerbicka, pointed out that automated model validation is one important research area of online simulation. It is a grand challenge to realize the concept of a fully automated model validation in a semiconductor manufacturing environment. The complexity and dynamics in semiconductor manufacturing are massive. It is also highly useful to identify and to correct model errors in all phases of the model creation process.

Due to the high complexity and the large problem size, the pure automated model validation is not in the range of interest of this thesis. Still, several aspects of model validation are addressed. For this online simulation project, the model validation is based on three elements:

- Automated data input validation
- Automated and manual data correction
- Manual and semi-automated model validation of the simulation model

One major focus of this thesis is to handle the data quality, which includes the validation of the input data. The cause of numerous errors in the simulation model is the input data quality. Therefore the objective is to identify and to fix those errors as early as possible. It is necessary to handle errors early in the data model and not later in the simulation model. This thesis addresses data quality issues in the data input modeling chapter.

A multitude of possibilities exists for a future implementation of a fully automated model validation. The experience coming out of this project is that the following three validation elements are essential for online simulation:

- A detailed automated comparison of simulation results and real world results is useful to monitor and to improve the quality of future simulation models. Therefore it is necessary to wait until results from reality are available to compare them with the simulation forecast results.
- The automated comparison of two or more simulation models figures out differences between these models. The comparison of the model results informs the user about changes of the forecast due to fab incidents. The comparison of the model itself identifies those incidents. Therefore it provides the reasons, why the forecast is changing.
- A unit test of the current simulation model provides an abstract evaluation of this model. The advantage is that it is not necessary to wait until the results from reality are available. Test cases of the model itself provide basic information whether the model generation is successful or not. Test cases for the simulation model results inform the user, whether the model behavior is within expectations or whether the model behavior alters, compared to the expected behavior.

6 Accuracy Results

This section contains the accuracy analysis of the online simulation results. The first parts represent the accuracy of forecasting results for different levels of detail. A major part is the analysis of the stochastic effects and the analysis of the forecast time horizon. The last part contains the sensitivity analysis of model initialization elements and their influence on simulation accuracy.

6.1 OverviewAccuracy Measurement

One challenge of this thesis is to describe the achieved accuracy of the online simulation forecast. It is important to measure the deviation from reality for a particular level of detail and the corresponding forecast horizon. It becomes necessary to analyze the accuracy for several key performance indicators (KPI) at fab level, work center level, and lot level.

Accuracy Measurement

It is a very complex task to measure the accuracy of the simulation forecast results. The reason is that different levels of detail exist (Law and Kelton 1999). Several KPIs exist as well (Hopp and Spearman 2000, SEMI 2003 a). Some KPI are suitable, other KPIs are not suitable for a particular level of detail. Also for each KPI multiple ways of computation exist. Furthermore it is necessary to differentiate a single run and confidence runs of the same model. It is also feasible to analyze multiple models at different points in time.

In literature there are also several forecast performance measures available (Makridakis et al. 1998, Hyndman and Koehler 2005). The interpretation of the results depends on the particular forecast performance parameter and its strengths and weaknesses. To measure the deviation from reality the mean absolute error (MAE) and the mean absolute percentage error (MAPE) are used because these measures are simple and intuitive. The formulas below describe the basic way of computing the accuracy. The value SIM$_i$ represents the simulation based forecast value. REAL$_i$ is the observed value from the real wafer fab.

$$MAE = \frac{1}{n} \overset{n}{\underset{i=1}{}} \quad SIM_i - REAL_i$$

$$MAPE = \frac{1}{n} \overset{n}{\underset{i=1}{}} \quad \frac{SIM_i - REAL_i}{REAL_i} * 100$$

Experimental Design

Table 32 contains an overview of the experimental designs to portrait the forecast accuracy. Three levels of detail exist. The fab level is the lowest level of detail. The lot level is the highest level of detail. The work center level is the desired level of detail for the short term forecast. For each level of detail, a single run generates the typical forecast results. The analysis results for multiple simulation runs from different points in time are available. It shows either trend changes or a stable behavior of the forecast parameter. Simulation runs for a different time horizon show trend changes over time. Further details about each experiment, the related parameter, and the results, are available in the corresponding section.

Run type	Fab level	Work center level	Lot level
Single run for 7 days forecast horizon	Moves/WIP	Arrival/WIP	Cycle time, matching%
Multiple runs at different point in time	Moves/WIP	Arrival/WIP	Cycle time, matching%
Time horizon with 1 day up to 30 days	Moves/WIP	Arrival/WIP	Cycle time, matching%
Stochastic effects by using confidence runs	Moves/WIP	Arrival/WIP	Cycle time, matching%

Table 32: Overview experimental design for accuracy analysis

Stochastic Effects

To analyze the stochastic effects in the simulation model, the results of the confidence runs are available for each level of detail. During the concept phase of online simulation, no decision was made whether a stochastic or deterministic mode modeling approach is used. The reason is that there is not enough information available. During model validation the question comes up, if the reason for the deviation from reality is the use of a single simulation run and not the use of an average value of the confidence runs. This thesis contains several experimental results to estimate the benefit of the confidence runs. The question is if the increase of accuracy is high enough, to use the average result from many confidence runs instead of the result of a single run. Therefore it is necessary to figure out which parameter is affected by stochastics and which parameter is not. It is important to quantify the total impact of the stochastic effects and to quantify the impact over time.

6.2 Fab Level Results

Fab level results have the lowest level of detail. The expectation is that the results are close to reality. The common opinion is that the detailed results on work center level are not valid, if the fab level results do not reflect reality.

On fab level the most important KPI are the fab moves and the fab WIP. The fab WIP is the sum of all lots in the fab. The fab moves consist of all lot moves for each piece of equipment in the fab. These parameters represent the fab behavior in simulation and reality. Other parameters like cycle time, wafer out, flow factor, and fab utilization also exist. For this online simulation model it is not useful to compare them, due to a different underlying definition. The fab cycle time and the wafer out are not part of this analysis, because of intermediate lot storage. Multiple product-dependent release and finish points exist, which are hard to compare. The risk of computation errors is high. Another reason is that the lot cycle time of finished lots highly depends on the initialization value, where the influence of the simulation run with only a few days is rather small. The flow factor and the average equipment utilization are also not comparable due to a different event structure in simulation and reality. One example is the internal queue time of equipment. In reality this time is part of the productive state. In simulation it is part of a non-productive state.

6.2.1 Single Run

The fab level forecast contains the WIP and the wafer moves results for a single model and a single simulation run with a seven days forecast period. Figure 83 and 84 depict the WIP and the moves per day for such a forecast. The scaling of the Y-axis starts at zero.

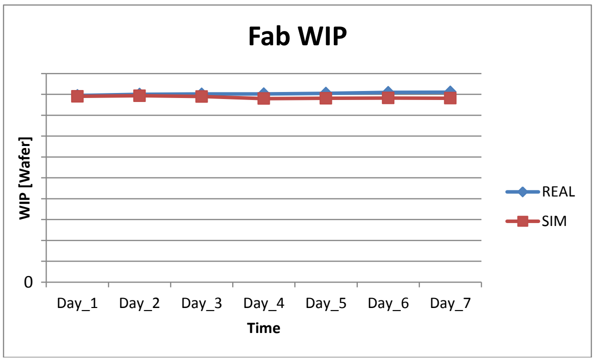


Figure 83: Fab WIP

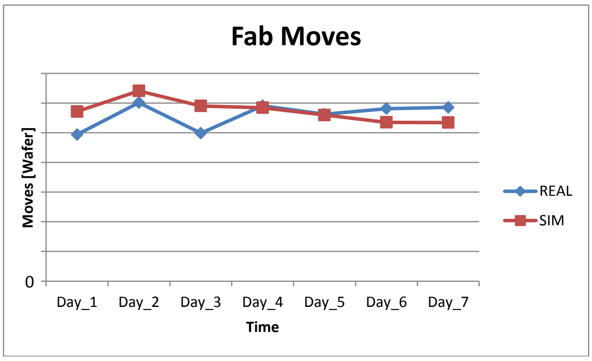


Figure 84: Fab Moves

In Figure 83 and 84, it becomes evident, that the fab level results reflect reality. Looking a little bit closer, the results are slightly too optimistic. The reason is that in reality many small disturbances exist in the manufacturing processes. It is not feasible to model all disturbances. Examples are unpredictable test wafer release, operator behavior, variability for transportation times, failures in the transport system, fab shutdowns, and irregular manual wafer inspection steps. Randell and Bolmsjö (2001) already pointed out, that a model generated out of fab data "..has limited stochastic behavior due to the lack of data for failures and cycle time variations in the ERP database". The presented online simulation model reduces the effect by using historical analysis. It generates statistics for several types of exceptions like equipment downs, hold, and rework.

6.2.2 Multiple Runs

To analyze the typical deviation between simulation and reality, it is further necessary to compute the forecast error for **multiple models** with a seven days forecast period.

Figure 85 and Figure 86 show the average deviation per day, between simulation and reality. The mean average percent error (MAPE) represents the deviation for fab WIP and fab moves. It has been generated for multiple simulation models, starting at different points of time. For example the forecast error in Figure 85, reaches a level of 2 percent, for a forecast horizon of 3 days.

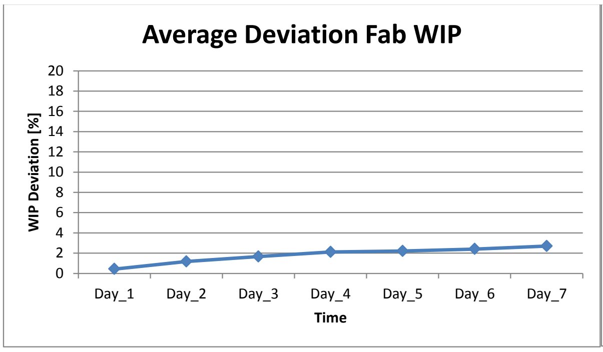


Figure 85: MAPE for fab WIP

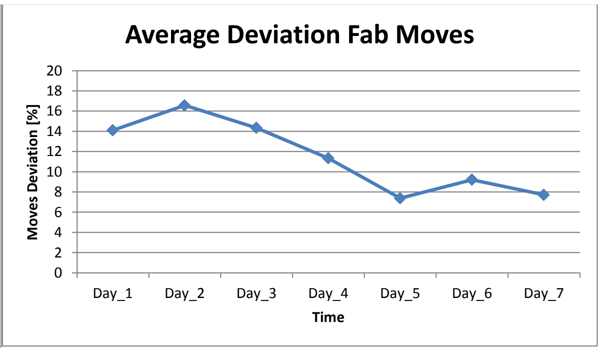


Figure 86: MAPE for fab moves

In Figure 85 and 86 the deviation (MAPE) for a seven day forecast is about 2 percent for fab WIP and about 11 percent for fab moves. The fab WIP is closer to reality than the fab moves. The reason is that the model is slightly too optimistic, as discussed before. It is interesting to see, that the deviation of the fab moves decreases over time. The expectation is that the deviation from reality increases, as seen for the fab WIP. A possible reason is that between simulation and reality small work center throughput differences exist. After the model initialization it takes some time until a stable behavior of the fab model is reached.

6.2.3 Time Horizon

The effect of an increased forecast time horizon becomes interesting, to analyze whether the model behavior is stable. The results for a single model with a single simulation run and an **increased forecast horizon** of seven days are available. To increase the time horizon of the model, the only manual change in the model is the extension of the lot release plan. To do so, the current lot release plan has been replicated for up to four weeks. For the forecast, no future information is used. Figure 87 and 88 depict the absolute value of the fab WIP and the fab moves for up to 30 days.

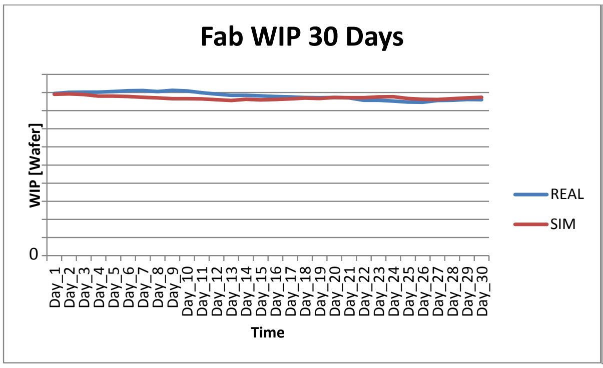


Figure 87: Fab WIP for 30 days

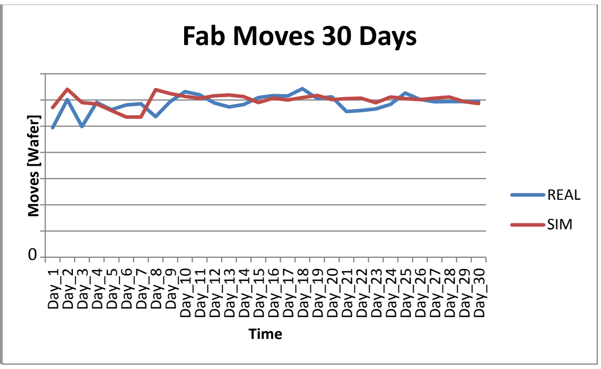


Figure 88: Fab moves for 30 days

The results on fab level reflect reality even for a very long period of time. The WIP and the moves results for reality show a stable behavior. In simulation the fab level results show a stable behavior too. The simulation model results on fab level are valid even after the targeted forecast horizon.

Figure 89 and 90 depict the deviation between simulation and reality for a 30 days simulation scenario. The daily MAPE is available for fab WIP and fab moves.

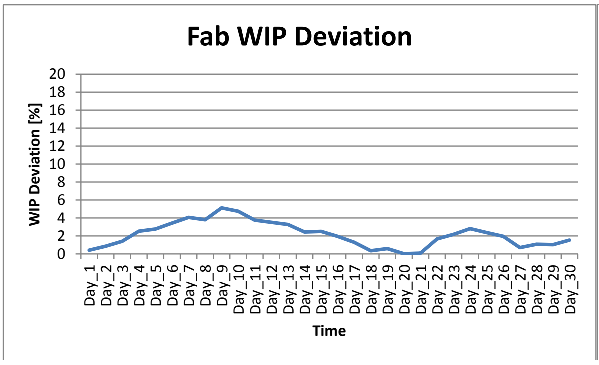


Figure 89: MAPE fab WIP for 30 days

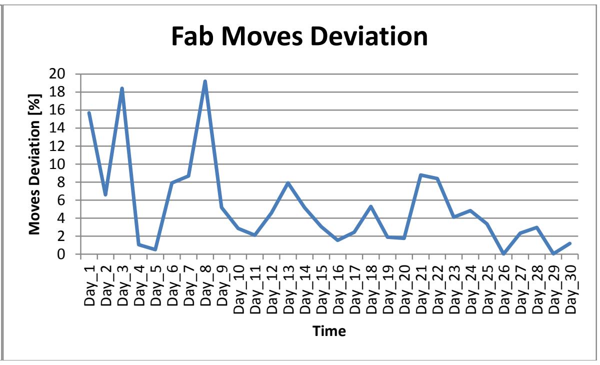


Figure 90: MAPE fab moves for 30 days

In Figure 89 and 90 it becomes clear, that the deviation from reality is very small. For the time period of 30 days, the average fab WIP deviation is also around 2% and the deviation for fab moves is around 5%. It is further interesting to see that for a single run the deviation does not always increase. The deviation for a single run is alternating, especially for the fab moves.

The main purpose of online simulation is the short term work center forecast and not the long term fab level forecast. To evaluate the long term fab level forecast accuracy, it is necessary to compare more than one sample over a very long period of time. For a long time period ramp up or ramp down phases also exist, depending on the current situation of the global market. To obtain reliable results about 100 simulation models from different points in time over a period of 5 up to 10 years are necessary. Another critical point is that the results for simulation highly depend on lot release, where the short term simulation includes no information about long term product mix changes. Long term equipment changes or fab shutdown incidents are also not part of the model. Therefore it is not feasible to expect an accurate prediction of major performance changes which exceed the forecast time period. Such a forecast and the comparison of the forecast quality are far beyond the scope of this thesis. The main result of the analysis is that this single simulation run shows a stable fab behavior. A permanent deviation is not visible. Slight fluctuations exist within the time horizon of 30 days.

6.2.4 Stochastic Effects

To analyze the stochastic effects, a single fab model has been executed with 10 confidence runs with different random numbers. The forecast horizon is 7 days. Figure 91 depicts the size of the confidence interval (SCI) in percent. The following formula computes the size of the confidence belt.

$$SCI_t = \frac{x_{t\,max} - x_{t\,min}}{\frac{1}{n} \; {}_{i=1}^{n} x_{ti}} * 100$$

$$x_{t\,min} = \text{Min } x_{ti}$$
$$x_{t\,max} = \text{Max } x_{ti}$$

Basically the formula displays the confidence interval as a difference between the minimum and the maximum value in percent, compared to the average forecast value. The value x_{ti} represents the forecast value for the day t and the confidence run i. The total number of confidence runs is n=10. The unit of the confidence interval is shown in percent.

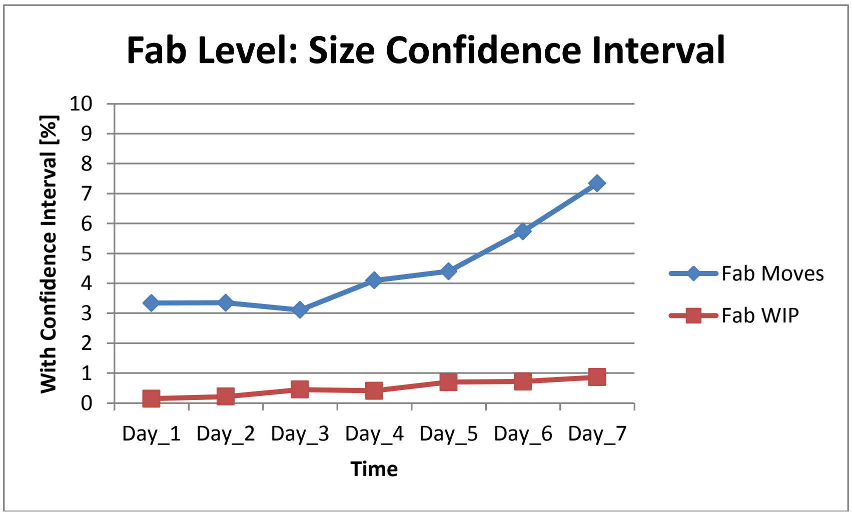


Figure 91: Fab level confidence interval

In Figure 91 the confidence interval for the Fab WIP is much lower than the interval for the fab moves. The moves are affected by stochastic effects, while the fab WIP is very stable. The stochastic effect on fab WIP is with less than 1% very low.

6.3 Work Center Results

The work center results are the primary results of the simulation based forecast. On work center level, several parameters like WIP, arrival, moves and cycle time exist. The main purpose is to predict trend changes for the work center performance. The lot arrival forecast is highly useful to predict trend changes. Another purpose is to alarm the production department if the situation becomes critical. Therefore the WIP forecast is effective, too. The work center lot arrival parameter and the work center WIP parameter have been selected for the accuracy analysis.

6.3.1 Single Run

For a single simulation run, the typical results for a single work center are depicted below. Figure 92 contains the work center lot arrival forecast. Figure 93 depicts the WIP forecast for the same work center. The corresponding data from reality is available to compare the quality of the forecast.

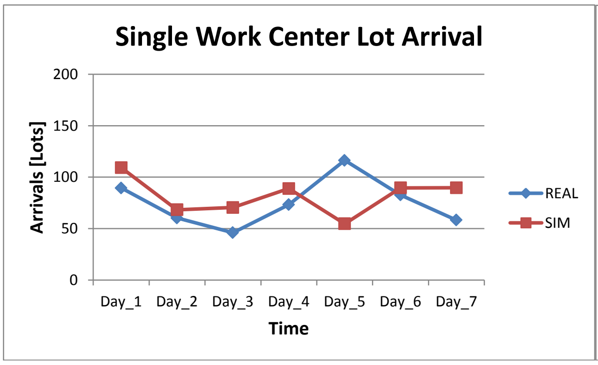


Figure 92: Work center arrival

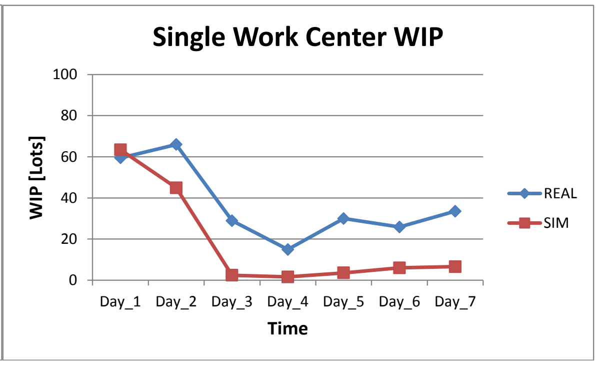


Figure 93: Work center WIP

The simulation results are close to reality. For this work center the simulation forecast is capable to predict trend changes. For the lot arrivals the forecast is excellent for the first 4 days. Only at the peak on day 5, the arrival forecast deviates from reality. Regarding the work center WIP, the prediction of the absolute value deviates from reality. Still the trend changes are correct. The WIP decreases until day 4. It slightly increases after day 4.

To analyze the accuracy on work center level, it is necessary to measure the deviation for all work centers. Figure 94 depicts the forecast error for the work center WIP. The black line in the figure depicts the mean absolute error (MAE) of all work centers per day. The red bars show the mean average percent error (MAPE). To avoid computation errors the zero values have been removed (Hyndman and Koehler 2005).

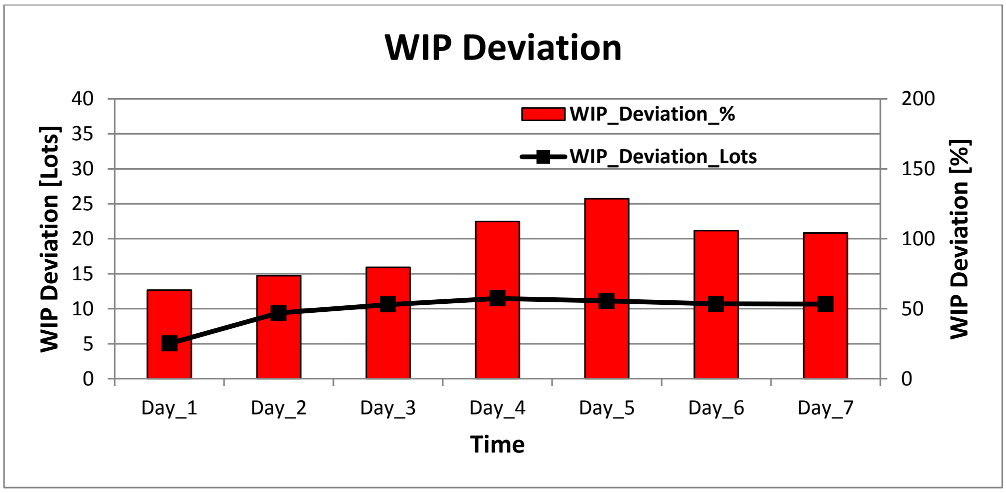


Figure 94: Work center WIP, MAE and MAPE

The absolute value (MAE) shows an average deviation of more than 10 lots per work center. The relative deviation (MAPE) reaches a value of more than 100%. The reason is that the average work center WIP is very low. In the whole wafer fab there are many small work centers. More than 100 work centers exist. They contain a small amount of equipment. So the

WIP lots are distributed to many work centers. Because of such a low average WIP, the MAPE is high. The conclusion is that for a valid work center WIP forecast, the accuracy is too low. Such a work center WIP forecast is not usable.

For the daily work center arrival deviation, the computation is similar to the work center WIP. The absolute error and relative error is depicted in Figure 95. The black line contains the MAE. The red bars show the MAPE.

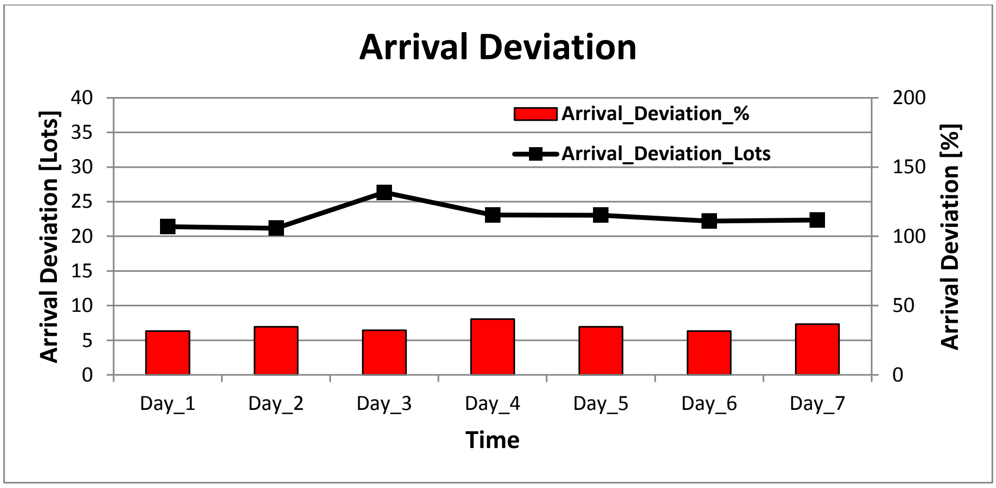


Figure 95: Work center arrival, MAE and MAPE

The absolute lot deviation between simulation and reality is about 23 lots. The relative deviation is about 40%. It is much better than the work center WIP. Still the number of 40% seems to be high, but it is useful to have closer look at it.

Figure 96 depicts the MAPE for the arrival deviation per work center, for the whole forecast period. This chart is sorted by the relative work center arrival deviation. Those work centers with the highest deviation of around 300% appear left. It is obvious, that the accuracy of the relative arrival forecast is very different in these work centers. In Figure 96, the work center moves are also available. The black line indicates how many percent of the fab moves are executed by those work centers. For example the peak for “Workcenter_070” with 3% indicates that this single work center executes 3% of the fab moves.

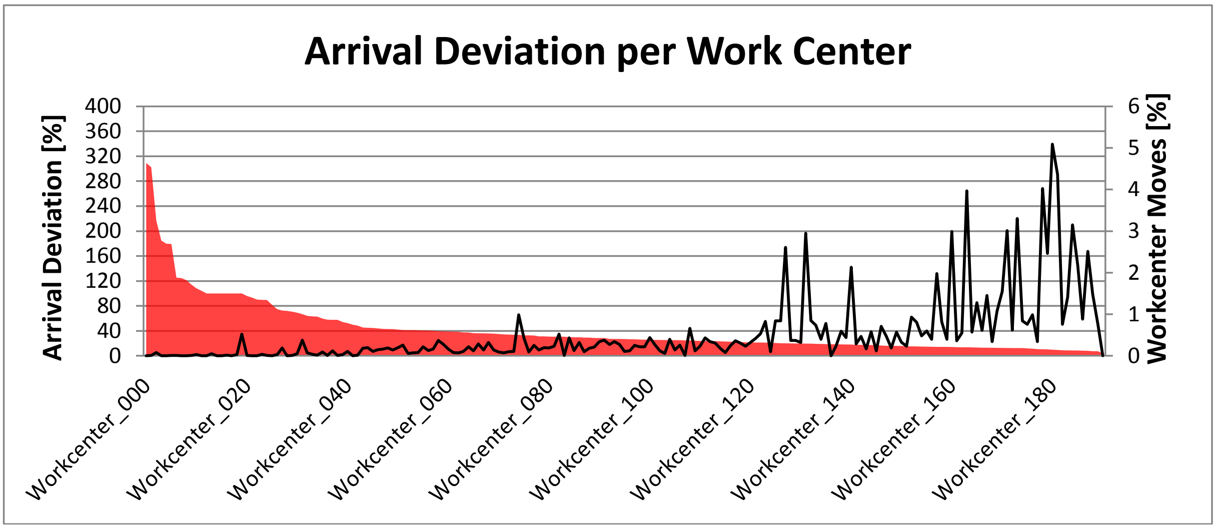


Figure 96: Arrival deviation MAPE per work center

For the work center arrivals, the MAPE is low and the results are accurate, if the work center moves are high. The relative deviation is high for those work centers with low moves. About one third of the work centers have a deviation that is higher than 40%. About one third of the work centers have a deviation between 40% and 20%. About one third of the work centers have a deviation below 20%, where the forecast is very useful. Table 33 lists these categories, including the relative number of work centers in the fab and the percentage of the fab moves for the whole category. Those work centers with the highest deviation above 40%, only carry out 4% of the fab moves. So to evaluate the accuracy, it is necessary to exclude work centers with low relevance from the accuracy analysis.

Work center category	Arrival deviation (MAPE)	Number of work center	Fab moves
1	40 % and above	30%	4.34 %
2	between 40 % and 20 %	37%	21.60 %
3	20 % or less	33%	74.06 %

Table 33: Work center categories for accuracy vs. fab moves

Work centers with a high number of moves are in the range of interest of the arrival forecast. The following figures contain the analysis for the lot arrivals MAPE. In contrast to the diagrams above only those work centers with a minimum of the fab moves are depicted. Figure 97 shows the relative arrival deviation (MAPE) for those work centers with at least 1% of the fab moves for each work center. It contains 11% of the work centers in the fab. Figure 98 shows the results for those work centers with at least 0.5% of the fab moves. It contains 26% of the work centers.

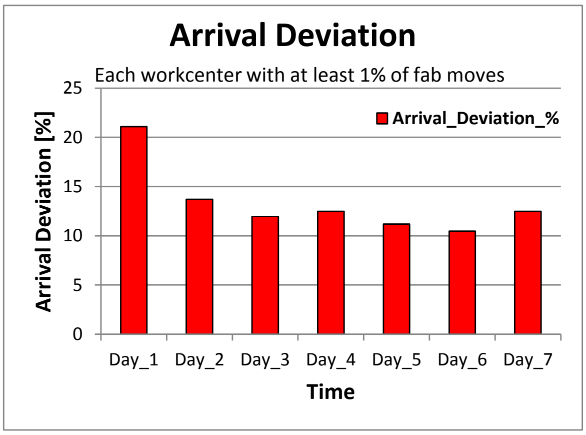


Figure 97: Work center, 1% of fab move

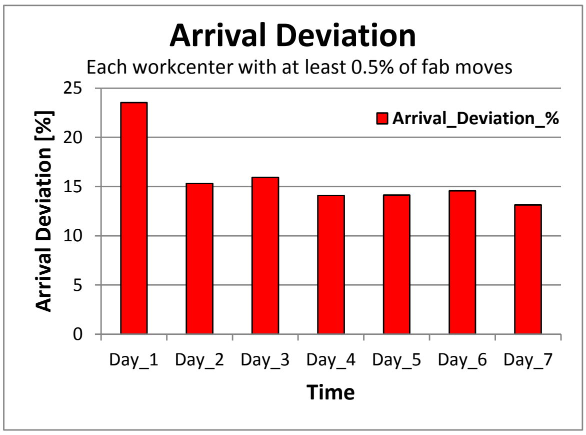


Figure 98: Work center, 0.5% of fab moves

For Figure 97, the relative lot arrival deviation (MAPE) reaches a level of about 13%. In Figure 98 it is 15%. The arrival deviation of the first day is much higher than for the rest of the time period for this work center subset. Despite the effort to initialize the fab properly, it still takes time to reach a stable behavior.

The conclusion from the analysis of a single simulation run is that the WIP deviation is high. It is not useful to forecast this value. However, the arrival forecast is close to reality. It is highly useful to forecast this value, for work centers with high number of moves. For this analysis only a single simulation run has been executed. In the next section, it is furthermore interesting to see whether the results from multiple simulation runs are in agreement with the results above.

6.3.2 Multiple Runs

To analyze multiple runs, the simulation based forecast results are available from different points in time. The first part contains the results for a single work center on four consecutive days. The second part contains the average deviation for all work centers for more than 10 different time periods.

For a single work center, multiple forecast results are available. Figure 99 depicts the same work center shown in the previous sections. Figure 100 , 101, and 102 show the forecast results for the work center for the next three days, too. The axis label for the time is fixed for all figures. For example day 4 is the same day in each figure. The results on day 4 with 73 lots arrivals in reality are the same. Only the simulation forecast results are different, depending on the particular time of the forecast.

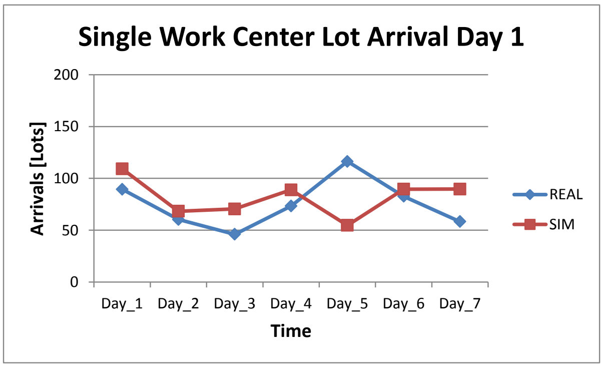


Figure 99: Forecast from day 1

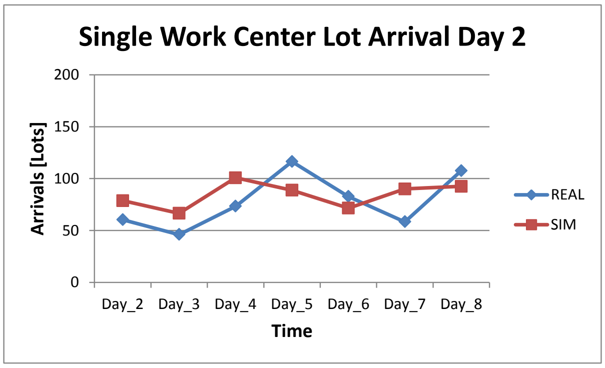


Figure 100: Forecast from day 2

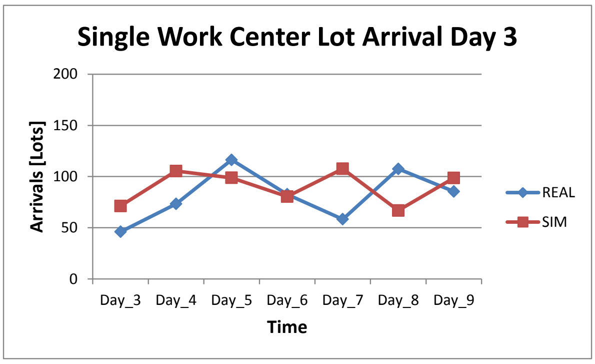


Figure 101: Forecast from day 3

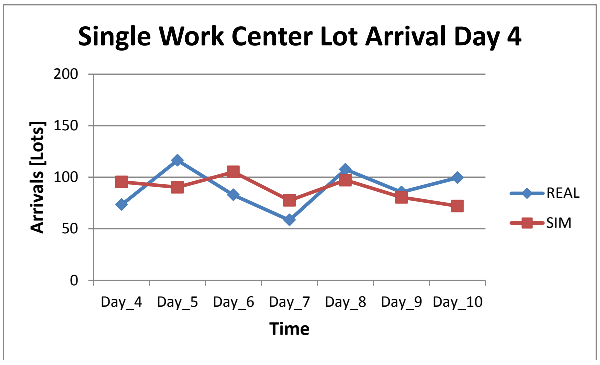


Figure 102: Forecast from day 4

In Figure 99 to 102 it can be seen that the forecast results are quite different. The first forecast in Figure 99 has good results except for the peak on day 5. The next forecasts in Figure 100 and 101 portray the peak on day 5 much better. Another peak is visible at day eight. The prediction on day two and day four is quite accurate (Figure 99 and 100), while on day three in Figure 101, the prediction of the peak is calculated one day too early.

The next step is to look at the average work center accuracy for multiple simulation models. Similar to the single simulation run, the absolute and the relative deviation (MAE and MAPE) is available for multiple runs, too. Figure 103 and 104 depict the work center WIP and the work center arrivals. For the time horizon, the label “Day_1” represents the average value for the first day after simulation start.

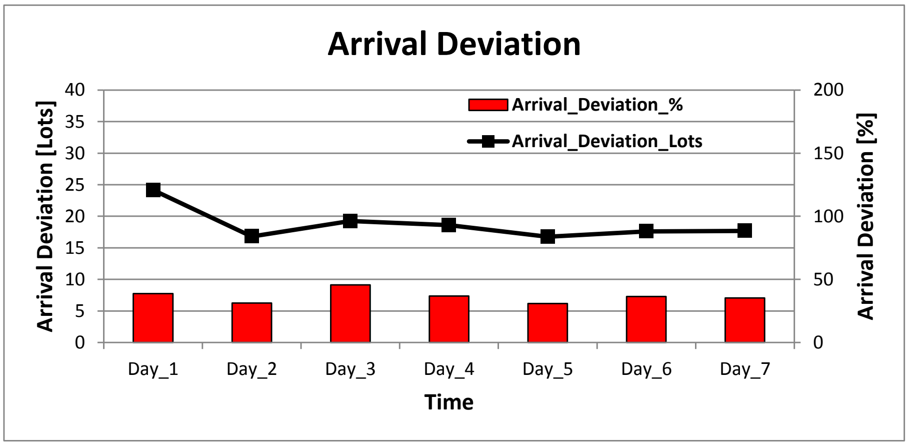


Figure 103: Work center arrival, absolute and relative deviation for multiple models

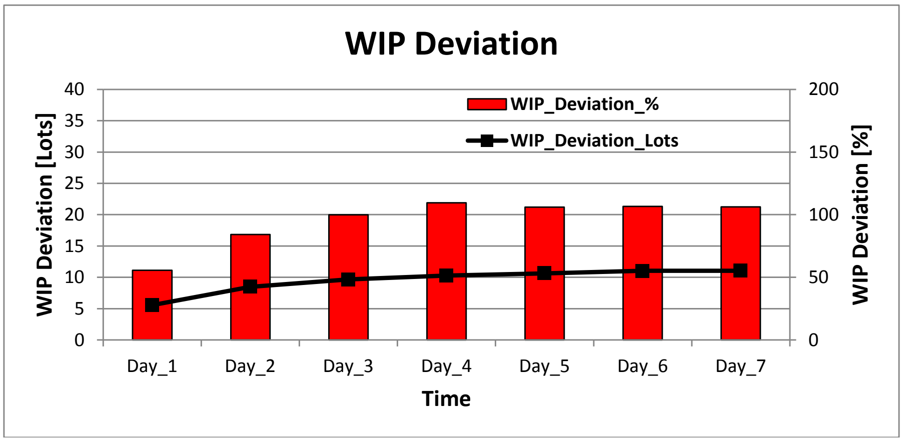


Figure 104: Work center WIP, absolute and relative deviation for multiple models

The relative arrival deviation for all work centers is very stable. It reaches a level of about 40%. The relative work center WIP deviation is very high. It is rising from 55% on the first day to 109% on day four. After one day, the WIP deviation remains stable. The results for multiple runs are very close to the result of a single run. Therefore a single simulation run represents the average behavior regarding the work center accuracy very well.

6.3.3 Time Horizon

The next part is the analysis of the time horizon and its effect on the work center accuracy. For this analysis, the forecast time horizon has been doubled from 7 days to 14 days. The simulation model has been validated for 7 days. The question is, whether the work center forecast deviation is stable or if it is going to increase. Therefore Figure 105 and 106 portray the average deviation (MAE and MAPE) for work center WIP and work center arrivals for a single simulation run.

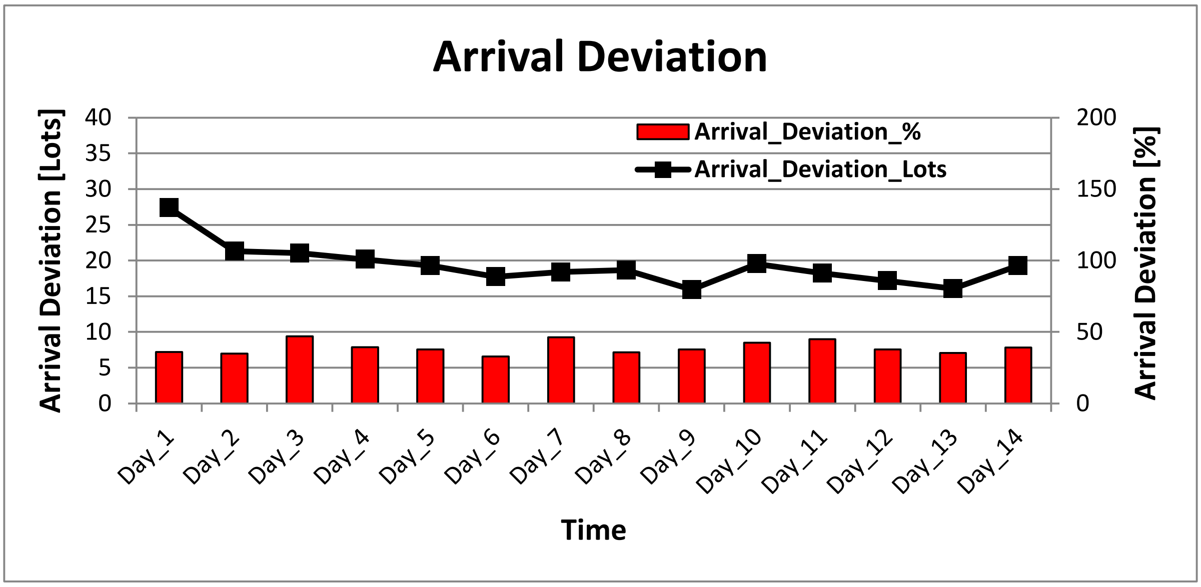


Figure 105: Work center arrival MAE and MAPE for 14 days

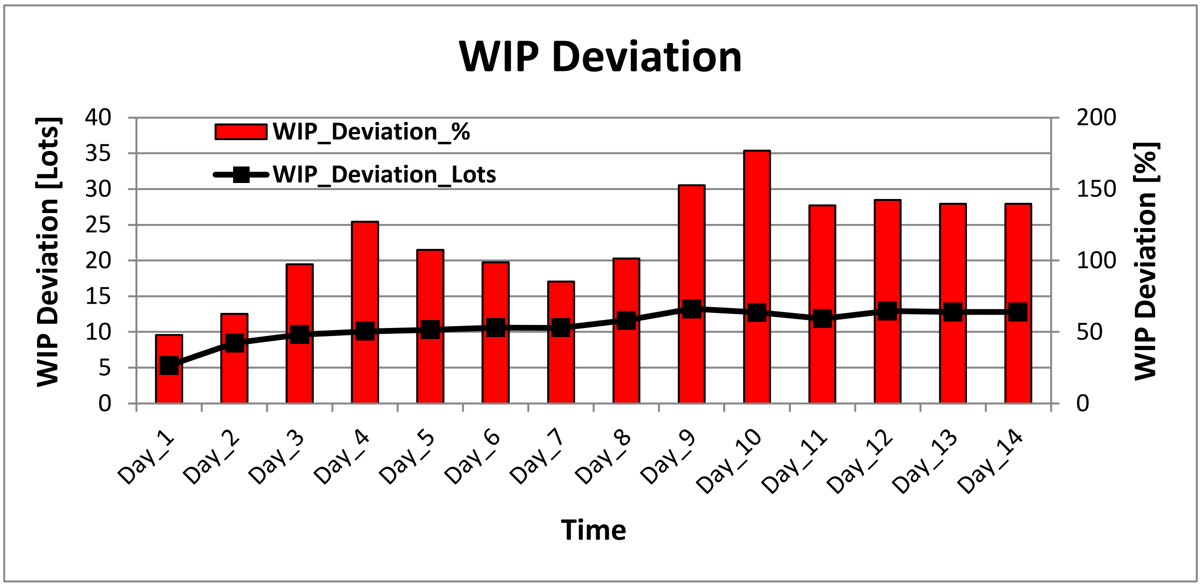


Figure 106: Work center WIP MAE and MAPE for 14 days

Figure 105 illustrates that the arrival deviation is relatively low and stable. Even for increased time horizon of 14 days the accuracy is quite stable. So it is feasible to increase the time horizon for the work center arrival forecast. The average work center WIP deviation in Figure 106 is already high at the beginning, it is fluctuating over time, and on average it is further increasing with expanded time horizon. Due to the low accuracy of the WIP it is not necessary to even analyze it any further.

It is further interesting to see, what a particular 14 days arrival forecast for a single work center looks like. As an example a single work center forecast is depicted in Figure 107, including the comparison to the real values.

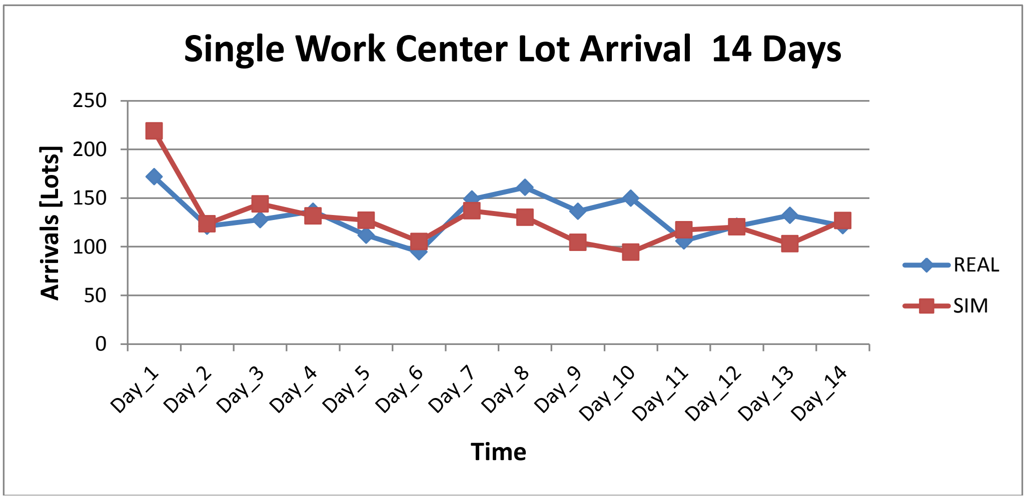


Figure 107: Work center arrival simulation, reality for 14 days

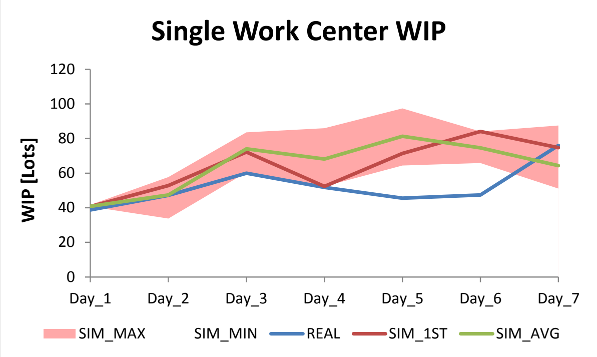
The average behavior for 14 days is quite stable. Slight fluctuations exist. It is clear that for this single run the prediction of the first 7 days is quite good. For the second week slight deviations arise. The peak from day 7 to day10 is lower in simulation. The peak on day 13 has been predicted on day 11 and 12, which is too early.

6.3.4 Stochastic Effects

Multiple confidence runs are carried out to show the stochastic effects on work center level. An example of the forecast results is available for a single work center. In addition, the average confidence interval of all work centers will be analyzed. The impact of the equipment downs as a major source of uncertainty will also be investigated. The last question is to figure out how much accuracy increase is possible by executing confidence runs.

Single Work center

Figure 108 and Figure 109 portray the WIP and the lot arrival forecast for a single work center. These charts contain forecast values for a single run “SIM_1ST" and the real work center performance “REAL”. In addition, 10 confidence runs have been executed to compute the average value “SIM_AVG”, the minimum “SIM_MIN”, and the maximum “SIM_MAX” of all confidence runs.



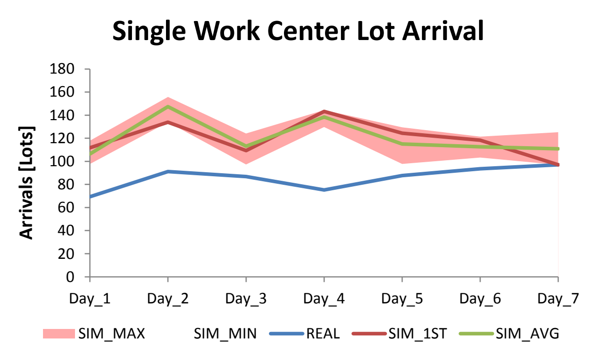


Figure 108: Work center WIP

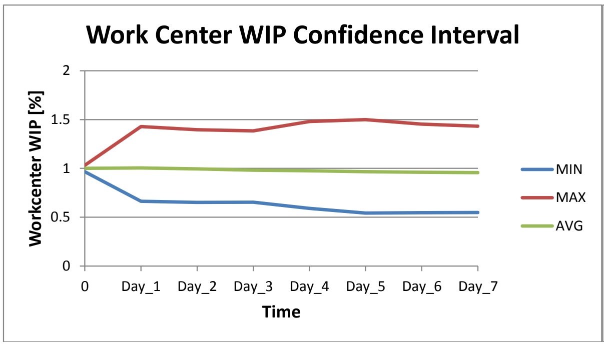
Figure 109: Work center lot arrival

This single example of a work center forecast shows the typical work center behavior. The results for a single run “SIM_1ST" are very close to average value of the confidence runs “SIM_AVG”. The single run results alternate around the average forecast value. The size of the confidence belt also differs much between several work centers, but also between the WIP and the lot arrival. For the work center WIP a deviation from reality appears on day 5 and day

6. The real value is not within the confidence interval. For the lot arrival the deviation from reality is also high and not within the confidence belt at this work center. The number of lot arrivals in simulation is higher than in reality. It is obvious that the reasons for the deviation between reality and simulation are not only caused by variability. Systematic reasons, like data quality issues are responsible, and not variability. It becomes necessary to increase the data quality. Data validation and model validation is an ongoing task to increase the quality of the results to keep it on a high level.

All Work center

To evaluate the stochastic effects on work center level, it is necessary to evaluate the stochastic effects not only for a single work center. Therefore the following figures depict the confidence interval and the average value for all work centers. Figure 110 displays the work center WIP. Figure 111 shows the work center arrival results. To compute the confidence interval, 10 confidence runs have been executed. Due to the non-disclosure agreement, the absolute scaling in the figure is hidden. It is available as a percentage value. The average value on the first days is 100% by definition.



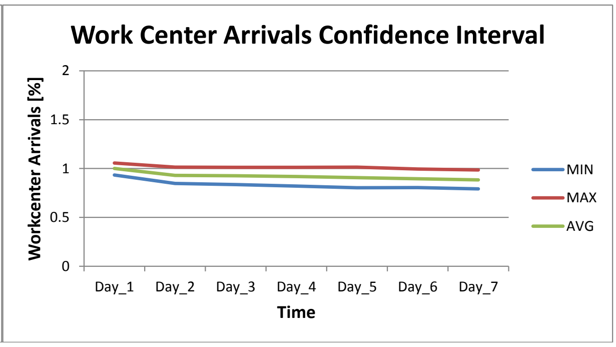


Figure 110: Conf. belt work center WIP

Figure 111: Conf. belt work center arrivals

In Figure 110 the size of the confidence interval for the work center WIP is very large. At time zero, which is the simulation start, the confidence interval is zero. After one day, it reaches a size of about 90% compared to the average WIP. The size of the confidence belt remains quite stable until the end of the last day. It becomes evident that the stochastic behavior of the wafer fab highly affects the work center WIP. For the lot arrivals in Figure 111, the confidence interval is much smaller. The size of the confidence interval between the minimum value and the maximum value is less than 20%, compared to the average number of lot arrivals. So for the work center arrivals, the deviation from reality and the influence of stochastic effects is very small. The work center arrival results are very useful for a simulation forecast. A single simulation run is also capable to produce acceptable results, compared to time consuming confidence runs. In this case the gap from the average work center arrivals to the minimum and maximum value of the confidence belt is only +/- 10 percent.

Influence of Equipment Downs

Another experiment analyzes the width of the confidence belt for the work center WIP and for the arrival forecast. Multiple questions are in the range of interest. First it is interesting to figure out the width of the confidence belt for WIP and arrival. Another part is to figure out how big the effect of the equipment downs is, as a major source of variability in the wafer fab.

The following formula computes the size of the confidence belt. The value WIP_{tiw} is the WIP of work center w, on day t, for confidence run i. The value m is the number of all work centers. The average size of the work center confidence interval per day is SCI_WIP_t.

$$SCI\_WIP_t = \frac{1}{m} \sum_{w=1}^{m} (\mathrm{WIP}_{tw\ max} - \mathrm{WIP}_{tw\ min}) \qquad \begin{array}{l} \mathrm{WIP}_{tw\ min} = \mathrm{Min}\ \ \mathrm{WIP}_{tiw} \\ \mathrm{WIP}_{tw\ max} = \mathrm{Max}\ \ \mathrm{WIP}_{tiw} \end{array}$$

The width of the confidence belt is depicted in Figure 112 and in Figure 113. As expected, the confidence interval is increasing over time. Short after simulation start it is increasing a lot. After about 4 days it is already quite stable. To estimate the influence of equipment downs a second scenario has been simulated with real equipment downs. To obtain the data for real downs it is necessary to wait seven days until the equipment down trace is available in reality. For a real forecast, the values of future downs events cannot be used. For the random down scenario in Figure 112 the size of the confidence interval converges with around 17 lots. For the real down scenario the average size of the confidence interval for WIP is about 9 lots, which is almost half. The equipment downs highly affect the confidence interval. For lot arrivals, this effect is similar as it is for the WIP.

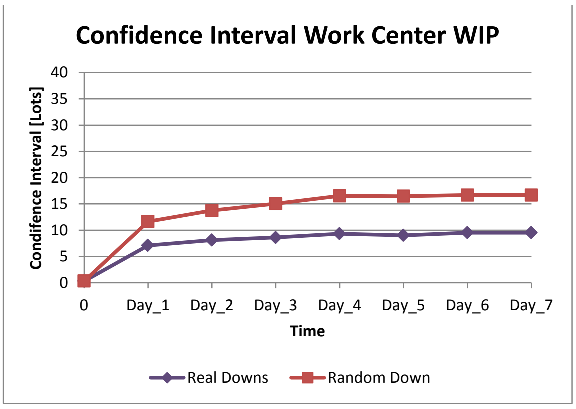


Figure 112: Conf. belt work center WIP

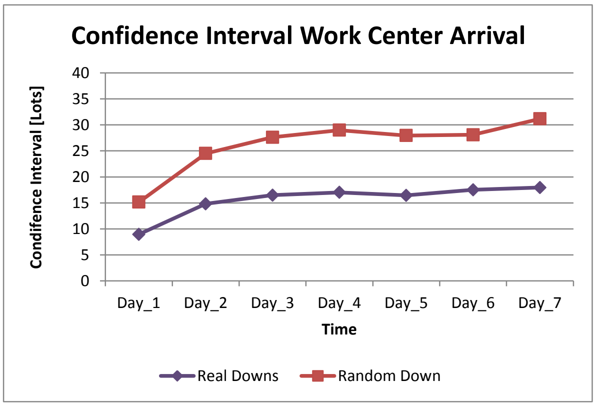


Figure 113: Conf. belt work center arrival

Figure 113 displays the size of the confidence belt with real downs and with random downs. On the first day after simulation start, the difference between the maximum and minimum is still small. The size of the confidence belt for lot arrival increases fast and converges with about 30 lots. With real equipment downs, the size of the confidence interval reduces to almost half of its value. The equipment downs are a major source of uncertainty.

Accuracy of a Single Run vs. Multiple Confidence Runs

Another question of this thesis is it to figure out if confidence runs increase the forecast accuracy for online simulation. Therefore it is necessary to compare the results accuracy from a single run and the average results over all confidence runs. The confidence belt is already very small for the work center arrivals. For the work center WIP the stochastic effects are significant. Therefore it is interesting to see the accuracy gain of multiple confidence runs. Figure 114 depicts the work center WIP deviation MAE for a single simulation run and the MAE of the average work center WIP deviation for 10 confidence runs.

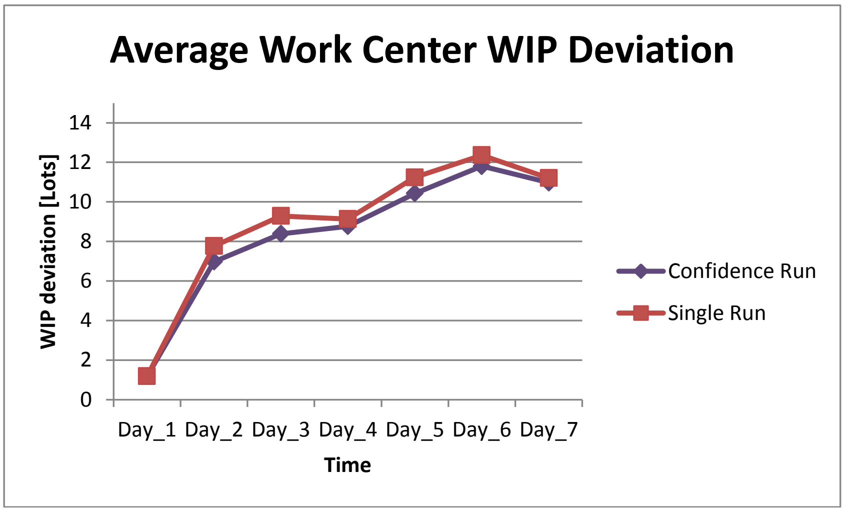


Figure 114: Average work center WIP deviation with single run and with confidence runs

For all work centers in the fab, the average results are almost the same executing multiple confidence runs or a single run. In Figure 114, the WIP deviation for a single run is only slightly higher than the WIP deviation for the confidence runs. It is also interesting to see the numbers behind. For a single run, the average WIP deviation is 8.8 lots. For the average value of the confidence runs, the average work center WIP deviation is 8.3. So the average work center WIP forecast is about 0.5 lots better, which is about 6%. The deviation from reality is reduced. So the use of confidence runs slightly increases the forecast accuracy. The disadvantage is the increase of the computation time. Instead of a single 6 minutes simulation run, it takes 1 hour to complete 10 confidence runs. This process time increase is not acceptable. So the tradeoff is to make one simulation run and accept the accuracy decrease of 6%.

6.4 Lot Level Results

The lot level has the highest level of detail in simulation. For such a high level of detail the expectation is that the forecast is not applicable due to the high deviation from reality. The complexity and uncertainty of the wafer fab is too high to expect perfect results on such a detailed level. An analysis is still useful to quantify the accuracy on lot level. It is necessary to provide the current state regarding the achieved accuracy. Another important point is to match the achieved accuracy with the requirements for certain applications.

The first step is to define the key performance indicators (KPI) to compare single lot results in simulation and reality. On lot level different KPIs exist, compared to the fab level or the work center level. Common parameters are WIP, cycle time, and moves. It is clear that the WIP is not useful, because the WIP of a single lot is always one. The lot moves of a single lot are countable, but there is no difference between a critical bottleneck move and a non-critical move. A more intuitive parameter is the cycle time, which also reflects the differences between time consuming bottleneck steps and fast non-bottleneck steps. This parameter also tracks the progress regarding the lot moves indirectly.

6.4.1 Single Run

To compare the lot cycle time, the approach from Section 5.5.1 is used. The matching deviation parameter and the cycle time deviation parameter are useful to measure the forecast error on lot level. Both parameters are presented in detail with the corresponding results from online simulation.

Cycle Time Deviation

The first parameter is the cycle time deviation. It compares the lot arrival event “i” between simulation and reality. The time “t_SIM_LOT_ARRIVAL$_i$” is the point in time in simulation and “t_REAL_LOT_ARRIVAL$_i$” is the point in time in reality. To compare both events, the combination of the lot name and the operation number has to be the same in simulation and in reality. If multiple occurrences are available for the same combination of lot name and operation number, then the algorithm selects the first element. A reason is, for example, lot rework. In addition to the MAE for cycle time, the MAPE value is also available.

To compute the relative cycle time deviation (MAPE) the absolute cycle time deviation (MAE) is divided by the duration between the simulation start time and the time of the corresponding real lot arrival event.

$$MAE\_CT = \frac{1}{n}\mathop{}\limits_{i=1}^{n} t\_SIM\_LOT\_ARRIVAL_i - t\_REAL\_LOT\_ARRIVAL_i$$

$$MAPE\_CT = \frac{1}{n}\mathop{}\limits_{i=1}^{n} \frac{t\_SIM\_LOT\_ARRIVAL_i - t\_REAL\_LOT\_ARRIVAL_i}{t\_REAL\_LOT\_ARRIVAL_i - t\_SIM\_START}$$

The forecast error for the cycle time deviation per day (MAE) is depicted in Figure 115. The corresponding percentage value (MAPE) is available in Figure 116.

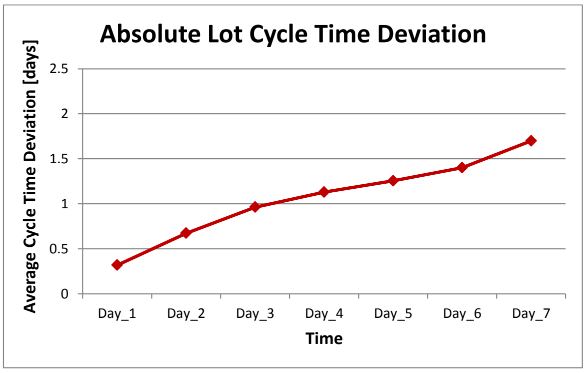


Figure 115: MAE lot cycle time

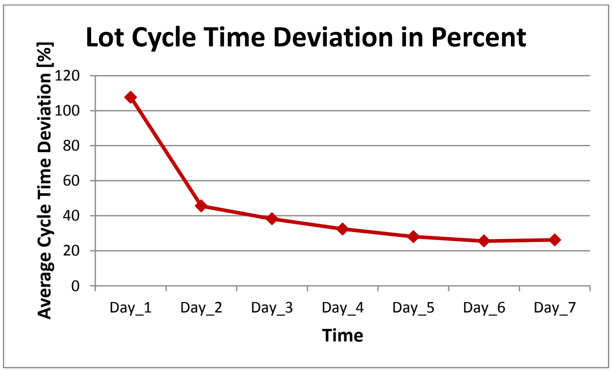


Figure 116: MAPE lot cycle time

In Figure 115 the absolute lot cycle time deviation increases. It starts at 0.4 days, after one day of simulation to a cycle time deviation of 1.7 days, after seven days of simulation. The trend of this value behaves as expected. In contrast, the cycle time deviation in percent in Figure 116 does not behave as expected. The deviation is very high at the beginning, and low at the end. It can be seen, that the cycle time deviation in Figure 116 also exceeds the level of 100 percent. The reason is that at simulation start, the cycle time in the denominator is close to zero which results in large numbers (Hyndman and Koehler 2005). The following example with anonymous real data is available to illustrate this behavior.

Lot	Route	Operation	Real Arrival [Min]	Sim Arrival [Min]	Abs Cycle Time Deviation [Min]	Abs Cycle Time Deviation [%]
Lot_02	Route_1	100	Before Sim Start	Before Sim Start		
Lot_02	Route_1	101	87	180	93	107
Lot_02	Route_1	102	248	275	27	11
Lot_02	Route_1	103	278	337	59	21

Table 34: Cycle time deviation for lot trace example from simulation and reality

Table 34 shows lot cycle data from the real lot trace and the simulation lot trace. Lot_01 arrives at operation 101 about 87 minutes after simulation start. In simulation it reaches this operation 180 minutes after simulation start. The difference is 93 minutes, which is about 107%, compared to the value in reality. For operation 102 and 103, the arrival times in simulation and in reality are grouped much closer. So it is clear that, especially at the very beginning of the simulation, the percentage value of the cycle time differences is high. This behavior is not intuitive. Still, the advantage of this forecast parameter is the high sensitivity at simulation start.

Matching Deviation

As mentioned before, another parameter is used to evaluate the forecast quality on lot level. This second parameter is the percentage of non-successful matching, see also Figure 117. This is the number of lot arrival events in reality, where no matching partner in simulation can be found, divided by the number of total lot arrivals in reality. The total number of lot arrivals in reality contains lot arrivals with successful matching events $i_{Matching}$ and with non-successful matching events $i_{Not\_Matching}$.

$$Matching_{Deviation} = \frac{i_{Not_{Matching}}}{i_{Not_{Matching}} + i_{Matching}} * 100$$

The purpose of the matching percentage is to evaluate the quality of the cycle time deviation. Without this number it is not clear how reliable the cycle time deviation is.

Figure 117 depicts the deviation of the matching percentage per day. For the first day, about 12% of the lot arrival events in reality do not have a corresponding lot arrival event in simulation.

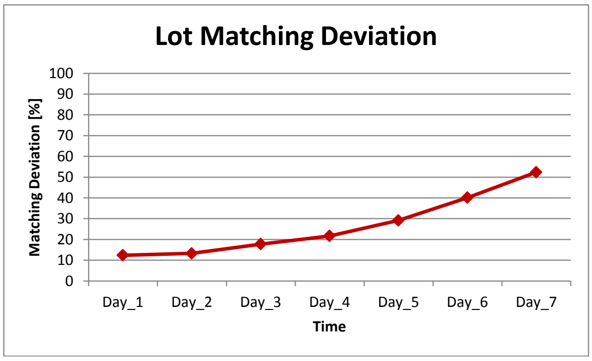


Figure 117: Lot matching deviation

The lot matching deviation increases to 52% on the last day. Several reasons exist why no matching partner is available in simulation. One reason is sampling. In reality a lot executes an operation, where in simulation it skips an operation due to sampling. In this case the corresponding lot arrival event is not available. Other reasons are rework, different lot names for split lots, and lot release sources without real lot names.

So it becomes clear, that the lot based cycle time comparison is very reliable at the beginning of the simulation. Only 12% of the lot arrival events do not have a corresponding simulation event. At the end of the simulation forecast horizon this kind of comparison is less reliable because 52% of the lots do not have a corresponding simulation event.

6.4.2 Multiple Runs

For multiple runs, the average results for the absolute cycle time deviation (MAE), cycle time deviation in percent (MAPE), and matching deviation are depicted in Figure 118 to 120.

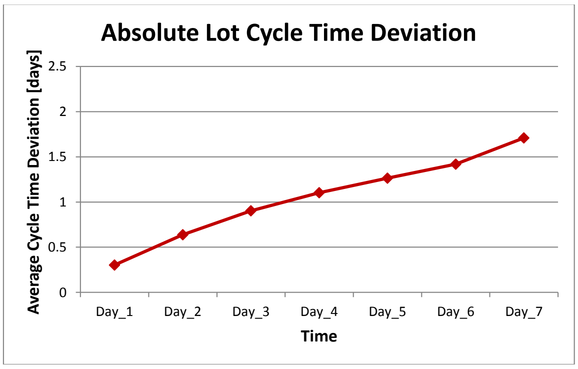


Figure 118: Absolute lot cycle time deviation

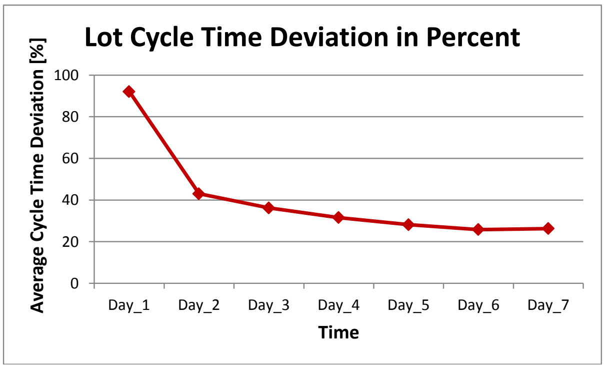


Figure 119: Lot cycle time deviation percent

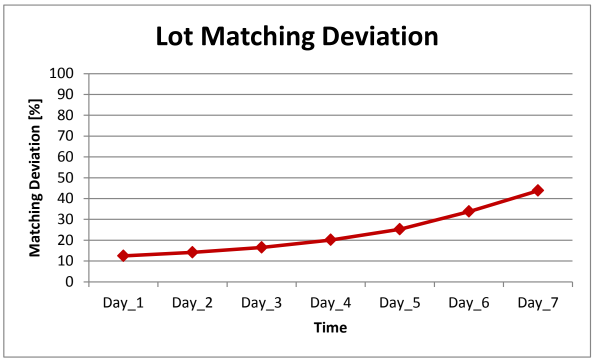


Figure 120: Lot matching deviation

The results of a single run and the average results for multiple runs are very similar. It becomes clear that the average results of multiple runs reflect the behavior of a single run.

6.4.3 Time Horizon

On lot level it is not useful to increase the time horizon even further because the lot matching of simulation and reality is done on basis of lot names. The lot release plan for the next 2 to 5 weeks does not have such a high level of detail. Future lot names are not available for this time period. The matching partners in simulation and reality cannot be identified.

For the time horizon analysis on lot level it is much more helpful to have a closer look at simulation start. Applications like lot scheduling need a lot arrival forecast. Such a forecast only captures a short time horizon. Therefore it is much more functional to quantify the deviation for the first couple of hours. For this analysis 10 confidence runs have been executed.

Figure 121 displays the cycle time deviation (MAE) on lot level, for the first 24 hours of simulation. Figure 122 displays the lot based matching deviation for the same time period. Figure 123 shows the minimum, the maximum, and the mean cycle time error on lot level (ME). The difference between the ME and the MAE computation is that the absolute element in the computation formula has been removed.

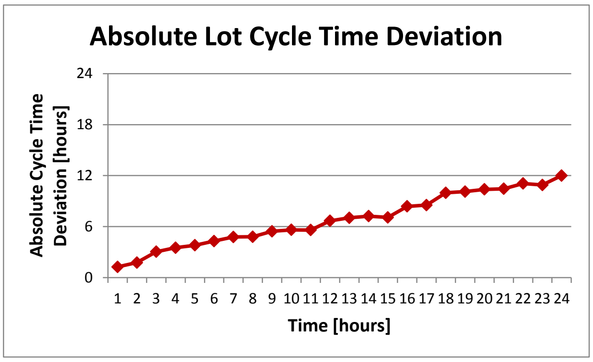


Figure 121: Cycle time deviation (MAE)

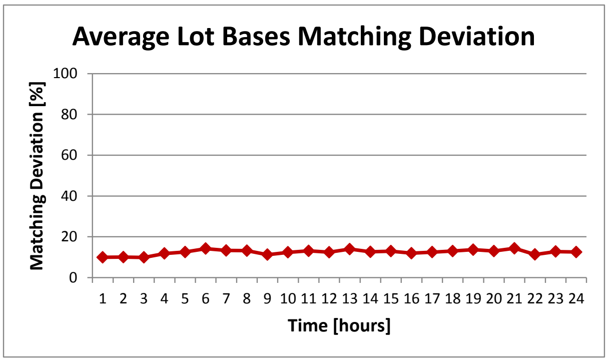


Figure 122: Avg. lot matching deviation

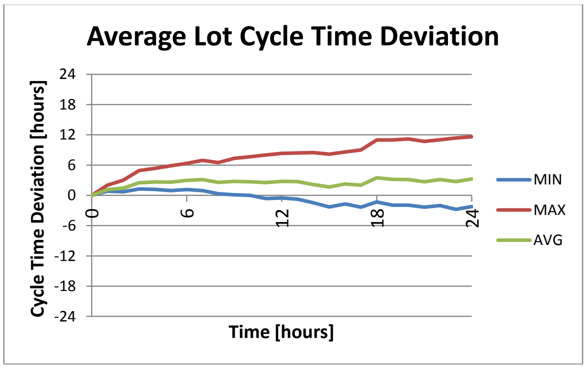


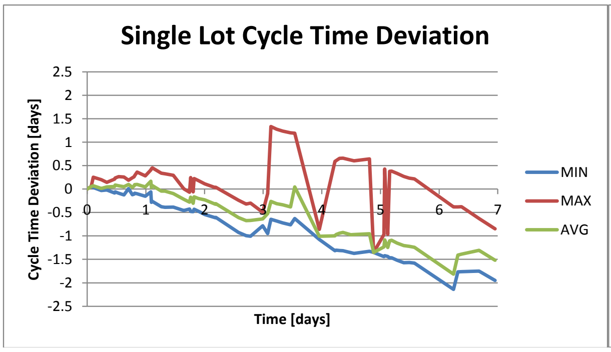
Figure 123: Cycle time error, minimum, maximum, and mean average

In Figure 121 the mean absolute error (MAE) for lot cycle time increases almost linear for the first 24 hours. After 24 hours of simulation, the mean absolute error reaches a level of about 12 hours. Regarding the matching percentage, the first 24 hour hours of simulation are very stable (Figure 122). On average 12% of the lots cannot be assigned, due to sampling rework etc. In Figure 123 after 6 hours, the mean cycle time deviation (ME) already reaches a deviation of 3 hours. After this time period, the size of the confidence interval reaches a size of 6 hours. For more than 6 hours the average deviation remains stable, while the confidence interval increases further.

6.4.4 Stochastic Effects

The lot level is also affected by stochastic effects, similar to the fab level and the work center level. 10 confidence runs have been executed to show the accuracy differences for simulation runs with different random numbers.

Figure 124 and 125 contain the minimum, the maximum, and the average cycle time deviation for a single lot. Two random lots have been selected to illustrate the stochastic effect for single lots. The minimum value contains the minimal cycle time deviation of all confidence runs for one point in time. The computation formula for the maximum value uses the maximum cycle time deviation for one point in time. The average value contains the average cycle time deviation of all confidence runs at one point in time.



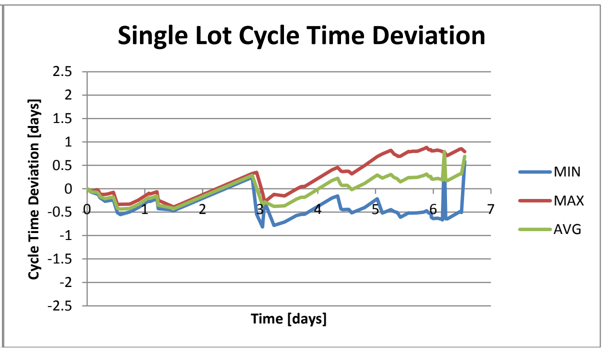


Figure 124: Cycle time deviation lot A

Figure 125: Cycle time deviation lot B

In Figure 124 the confidence interval of the cycle time deviation for lot “A”, reaches a level of +1.5 days down to -2 days within the first seven days forecast horizon. Lot “B” in Figure 125 has a smaller confidence interval of about +1 day, down to -1 day within the forecast horizon. It also becomes clear that for single lots, the confidence interval increases and decreases again. The average deviation also shows a systematic deviation which does not always depend on the particular random numbers. Especially for the first lot, the cycle time deviation becomes negative with around -1.5 days. It means the cycle time is much shorter in simulation than in reality. This lot is an example that the average fab behavior is too optimistic.

To see the full picture, it is also necessary to compute the confidence interval and the average value cycle time deviation (ME) for all lots. This average value does not contain absolute values. The purpose is to figure out if the average trend is positive or negative. This is an indicator if the fab model throughput behavior is too optimistic or too pessimistic. Figure 126 illustrates the average lot cycle time deviation and the confidence belt for 10 confidence runs, for a forecast horizon of seven days.

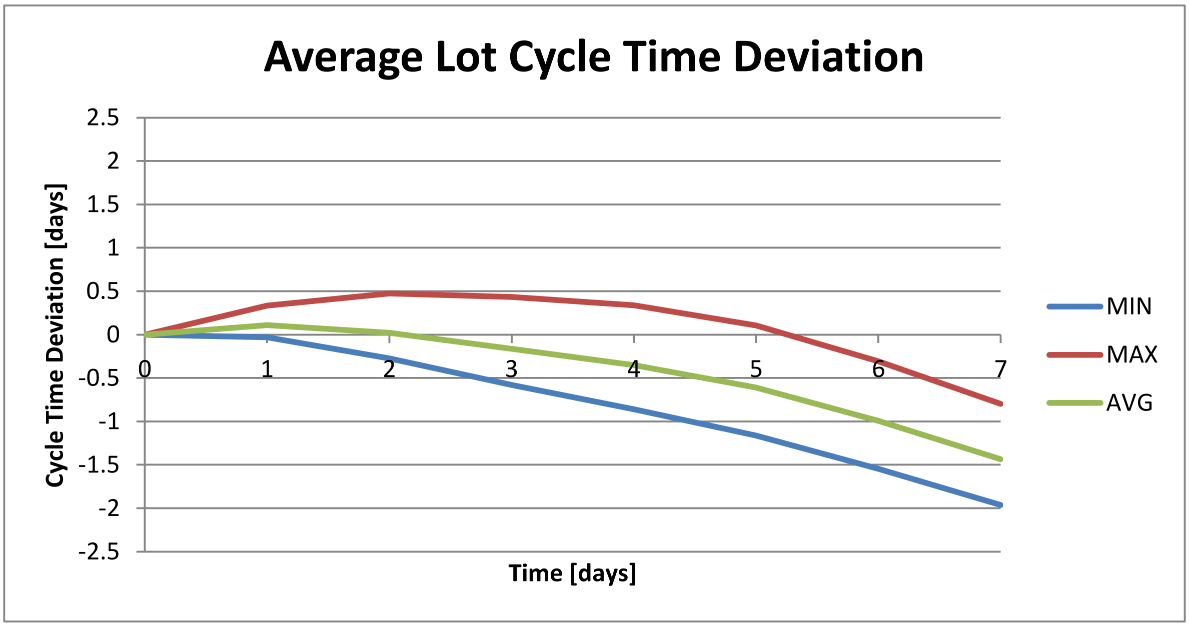


Figure 126: Average lot cycle time deviation with confidence belt

For the first two up to three days of simulation, the average cycle time deviation in Figure 126 is close to 0. At the end of the forecast horizon the average cycle time deviation shows a negative trend. The cycle time (CT) in simulation is smaller than the cycle time in reality. So, after six days of simulation, the lot has already gained one day cycle time on average. It becomes obvious that the fab model is too optimistic.

In addition to that systematic deviation, the stochastic behavior affects the accuracy even further. After 3 days of simulation, the size of the confidence interval, between the minimum and the maximum value, increases fast. It reaches a level of about 1 day. The size of the confidence belt with absolute values is also depicted in Figure 127. After day 4, the size of the confidence belt is quite stable. It increases to slightly more than 1 day. In addition Figure 128 depicts the relative size of the confidence interval. It has a size of about 37% for the first two days. Afterward it is decreasing to 17%.

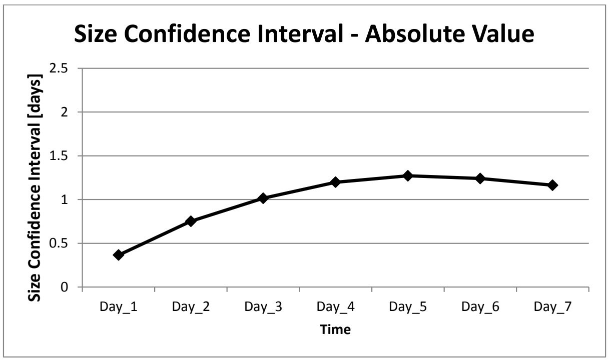


Figure 127: Confidence interval, absolute CT

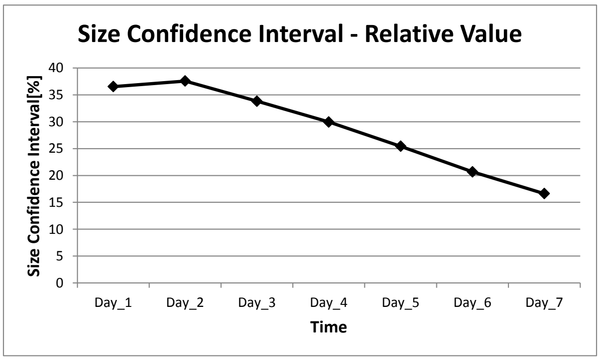


Figure 128: Confidence interval, relative CT

6.5 Sensitivity Analysis for Model Initialization

The modeling chapter contains an explanation what has been done to increase model accuracy. Examples are an increase of the level of detail for equipment modeling, attribute based sampling modeling, or detailed dispatch rule modeling. To increase the model accuracy, a major concept is the detailed model initialization. It is within the range of interest of this thesis to explore about the accuracy increase of a detailed model initialization. Another task is to figure out, if the detailed model initialization affects the accuracy for the whole forecast period under the aspect of the following ongoing statement: “If you cannot initialize fab, and predict the first day, how can you predict the future for two days onwards.” So it is interesting to observe changes in the increase of forecast accuracy on the first day, the second day, and so on. Several test cases have been executed to address these issues. The lot cycle time deviation has been selected to measure the accuracy increase, because this parameter is highly sensitive to measure the forecast accuracy.

6.5.1 Experimental Design

For this analysis several combinations of initialization elements have been executed. Figure 129 depicts these individual elements and their relations. For example to assign remaining process time, it is useful to occupy the current station first. This is also a constraint in the ASAP simulation software.

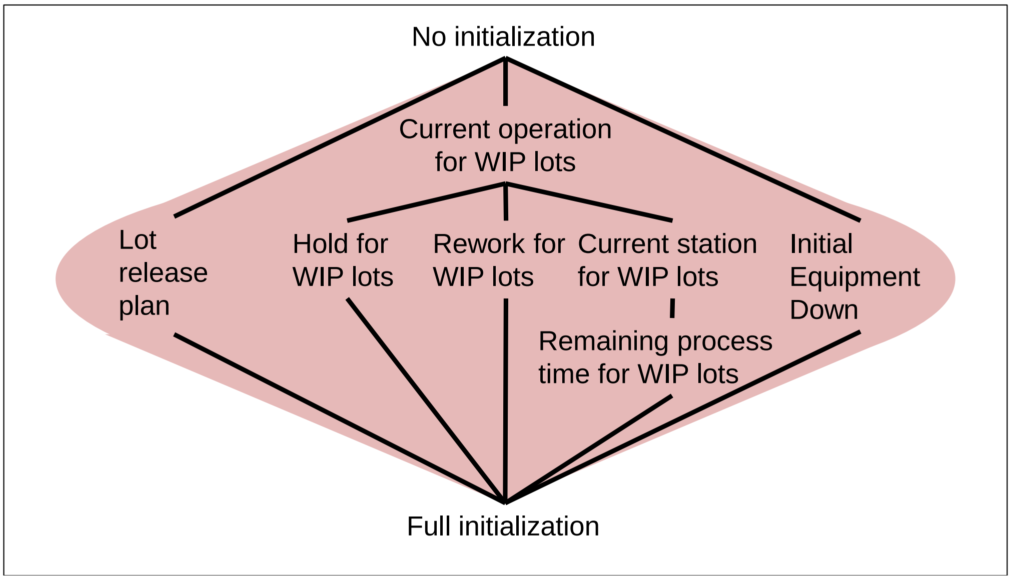


Figure 129: Elements for Initialization Scenarios

Table 35 shows the design of experiments. The left part compares very basic scenarios. It contains the scenarios with full initialization and without initialization. The purpose is to show the differences if the current operation of the lots is removed. The expectation is that the results do not match at all. This scenario is useful to obtain an impression about the accuracy impact for lot initialization with the current operation. The initialization with the lot release plan is also part of this accuracy analysis. The lot release distribution with the corresponding inter-arrival time replaces the lot release plan. The major difference is that without the lot release plan, the lot names for the future lot are no longer available to compare cycle times.

Basic Scenarios	Detailed Scenarios
1. No initialization 2. Initialization without lot release plan 3. Full initialization	1. Full initialization 2. Full initialization without current station and remaining process time 3. Full initialization without remaining process time 4. Full initialization without initial equipment downtime 5. Full initialization without a detailed lot release plan

Table 35: Design of experiments

The detailed scenarios are available in the right part of Table 35. The full initialization scenario is the reference scenario. The comparison of the impact of the single elements is in the range of interest. Therefore the single elements are removed and compared individually. For this analysis a single simulation model has been selected. Ten confidence runs have been executed for each scenario.

6.5.2 Results

The first part contains the results for the basic scenarios. For this comparison, the absolute cycle time deviation (MAE) is depicted in Figure 130. The matching deviation is available in Figure 131.

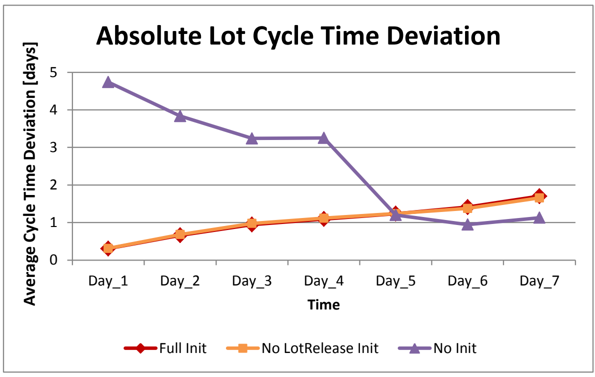


Figure 130: Lot cycle time deviation

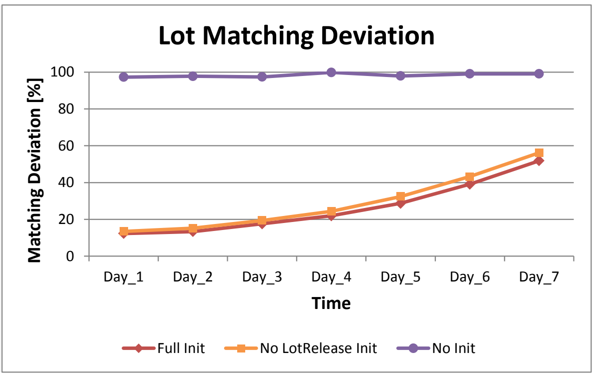


Figure 131: Lot matching deviation

For the scenario without any initialization, the cycle time error in Figure 130 is high. On average only 1.6% of the lots find a matching partner and 98.4% do not match. It shows that the impact of the lot initialization with the current operation is vast. As expected, the short term simulation results are not usable without the initialization of the current lot operation. The impact of the detailed lot release plan model initialization is very small. In Figure 130 no accuracy loss is visible, if the detailed lot release plan is not available. In Figure 131 the matching deviation is about 5% higher, if the detailed lot release plan is not available. Therefore it is useful to add the detailed lot release plan, but it is not imperatively necessary.

The second part contains the experimental results for the detailed scenarios. The reference is the full initialization scenario. The experimental design removes single initialization elements. The higher the accuracy loss of the simulation model is, the higher is the impact of the model initialization element. Figure 132 and Figure 133 compare the MEA and MAPE for the lot cycle time.

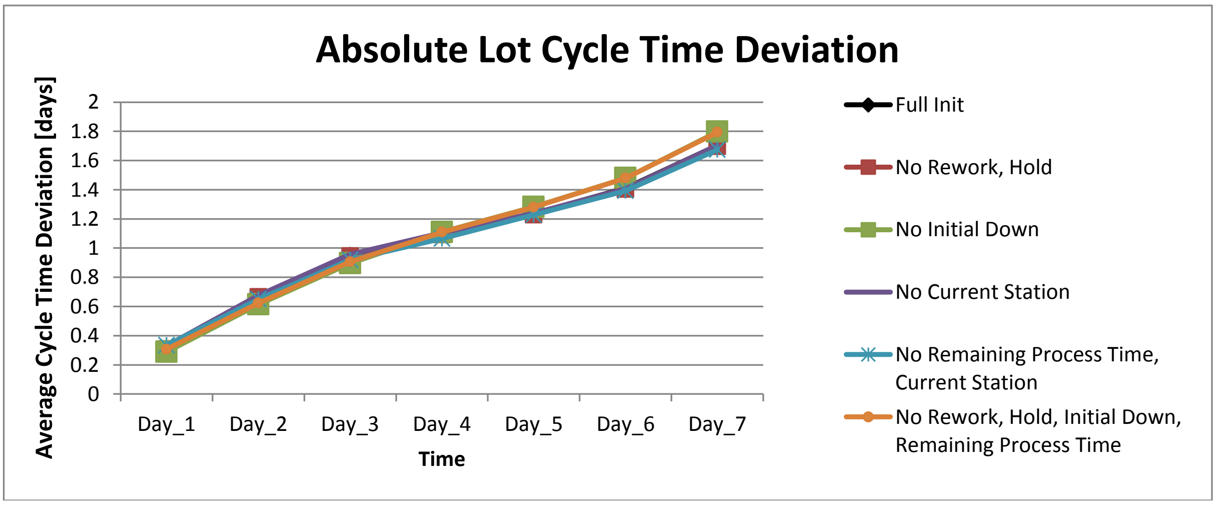


Figure 132: Mean absolute error (MAE) of lot cycle time for initialization scenarios

For Figure 132, the MAE results are very much alike from the first day onwards. Only for those scenarios where the initial downs have been removed, slight accuracy differences exist at the end of the forecast period. The absolute forecast error increases insignificantly. So even with a highly detailed initialization, the absolute forecast error is almost the same.

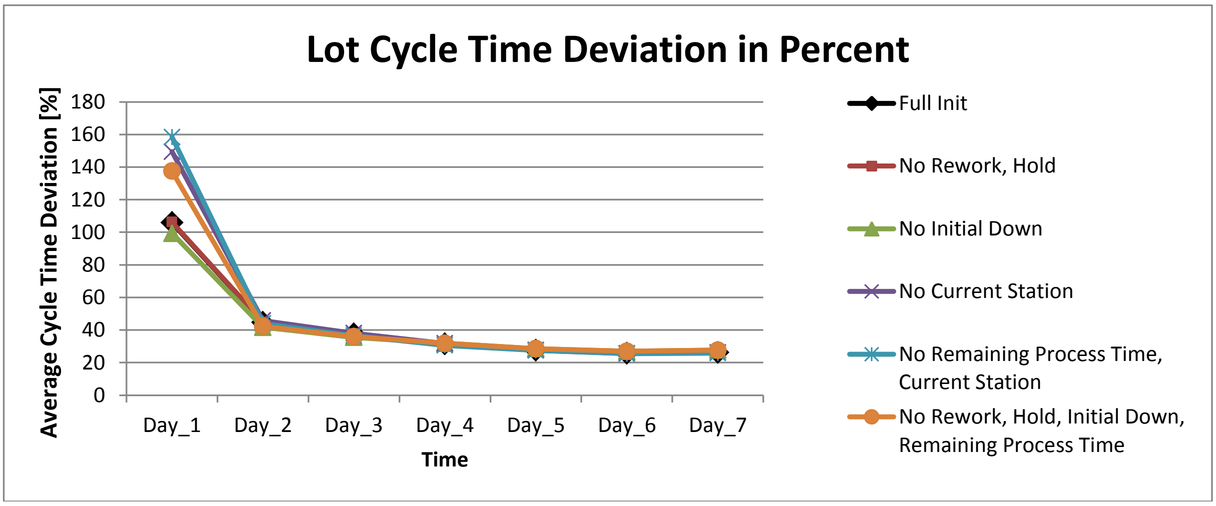


Figure 133: Mean absolute percent error (MAPE) of lot cycle time for initialization scenarios

In Figure 133 the MAPE for the lot cycle time is very sensitive for the results from the first day. High differences exist on the first day, whereby from day 2 onwards, the results are almost similar. For the analysis of the first day, basically two groups become visible. For the first group the cycle time forecast error is around 100%. This group contains the full initialization scenario and those scenarios where only the initial downs, the hold, and the rework initialization are removed. For the second group, the cycle time error on the first day is much higher, with around 150%. For those scenarios the current station has been removed.

Table 36 summarizes the effects of the model initialization. It becomes clear, that the initialization of rework and hold has almost no effect. To increase the accuracy for the first day, it is definitely useful to initialize the current station and the remaining process time. After one day of simulation, the accuracy advantage is lost. For model initialization, the equipment downs and the lot release plan have a long term effect on simulation accuracy. If the initialization of the equipment downs and the initialization of the lot release plan is not part of the model, than the forecast error slightly increases. Still, this effect is very small. As expected, the initialization of the WIP lots with the current operation is essential for online simulation.

Initialization element	Short term effect (1st day only)	Long term effect (2nd day onwards)
Hold, rework	N/A	N/A
Lot release plan	N/A	Small effect
Initial down	N/A	Small effect
Current station, Remaining process time	High effect	N/A
WIP lots with current operation	Essential	Essential

Table 36: Effect of the model initialization elements

6.6 Summary for Simulation Accuracy

This section summarizes the results of the accuracy analysis. The first part contains an overview of the sensitivity analysis. The second part reflects the accuracy results in a single diagram. This diagram combines the level of detail, the forecast horizon, and the forecast error. The third part summarizes the achieved accuracy for each level of detail separately.

The analysis of the stochastic effects describes the impact of uncertainties on different levels of detail. The last part discusses forecast accuracy results in the semiconductor manufacturing literature.

6.6.1 Sensitivity Analysis for Model Initialization

The analysis of the model initialization shows which model initialization elements are useful and which are not. The initialization of the WIP lots with the current operation is essential for short term simulation. This is a very basic model element. It shows the largest effect. The model initialization, with current station and remaining process time, increases the forecast accuracy only for the beginning of the forecast period. After the first day, the advantage is lost. Only the model initialization with the detailed lot release plan and the initial equipment downs has any positive effect on simulation accuracy from the second day onwards. The accuracy increase from these two elements is very small too.

So the question from the beginning was, if the model initialization affects the whole forecast period. The analysis shows, that a detailed model initialization, beyond the current operation of WIP lots, does not necessarily have a high impact on the accuracy of the whole forecast period. Either such an advantage is very small or is neglectable after the first day.

6.6.2 Relation Forecast Error, Time Horizon, and the Level of detail

One objective of this thesis is to characterize the accuracy of the simulation model forecast for different levels of detail. The previous sections describe the forecast accuracy in detail. Now the challenge is to provide an overview in one single chart.

For fab level, the daily WIP deviation has been selected. For work center level, the daily arrival deviation displays useful results. The cycle time deviation since simulation start represents the accuracy on lot level. These parameters have been selected, because they represent the accuracy best, for the particular levels of detail. As seen in the description of the previous sections, the following criteria are used to select these parameters:

- The values are comparable to reality
- The stochastic influence is low
- The KPI have the potential to be useful for a forecast

For each KPI, the average results from multiple simulations are used. These values represent the model deviation from different points in time with sufficient reliability.

Fab level

Days	1	2	3	4	5	6	7
WIP deviation [%]	0.43	1.16	1.66	2.11	2.20	2.39	2.69
WIP deviation [Lots]	10.89	20.86	35.63	64.40	68.97	76.17	82.94

Table 37: Fab level, average WIP deviation for multiple models

Work center level

Days	1	2	3	4	5	6	7
Arrival deviation [%]	38.82	31.36	45.69	36.88	30.92	36.45	35.32
Arrival deviation [Lots]	24.15	16.80	19.24	18.58	16.77	17.60	17.69

Table 38: Work center level, average arrival deviation for multiple models

Lot level

Days	1	2	3	4	5	6	7
Cycle time deviation [%]	92.07	43.02	36.22	31.58	28.14	25.81	26.29
Cycle time deviation [hours]	7.26	15.31	21.66	26.46	30.32	34.03	41.01

Table 39: Lot level, average cycle time deviation for multiple models

Figure 134 depicts the MAPE to provide an accuracy overview of the simulation forecast. The diagram shows the relations between the forecast deviation, the forecast horizon, and the KPI of the corresponding level of detail.

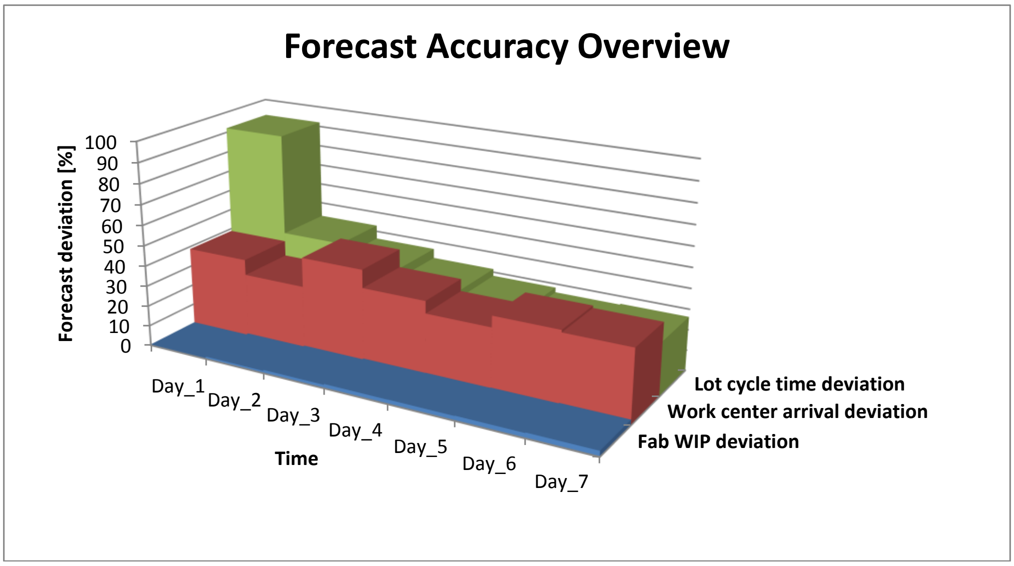


Figure 134: Forecast accuracy overview

In Figure 134, the relative deviation is very low on fab level, medium on work center level and high on lot level. Regarding the trend line, the relative deviation on fab level increases, even if the value is very small. The deviation is stable on work center level. On lot level, the deviation is decreasing. Such a decreasing accuracy is not expected. But it becomes explainable when looking at the computation formula for the MAPE of the lot cycle time. The relative cycle time deviation is defined by the time difference between lot arrival in simulation and reality, divided by the duration between the lot arrival time in reality and the simulation start time. The shorter the duration between a future point in time and simulation start, the smaller is the duration in the denominator, the higher is the relative deviation for lot cycle time (Hyndman and Koehler 2005).

This kind of result is not intuitive. Therefore the corresponding absolute forecast error (MAE) is depicted in Figure 135. The order of the z-axis has been changed, because the expectation is that absolute deviation becomes higher on fab level and lower for single lots. The figure shows that the deviation for work center arrivals is almost constant. For fab WIP and lot cycle time, the deviation increases.

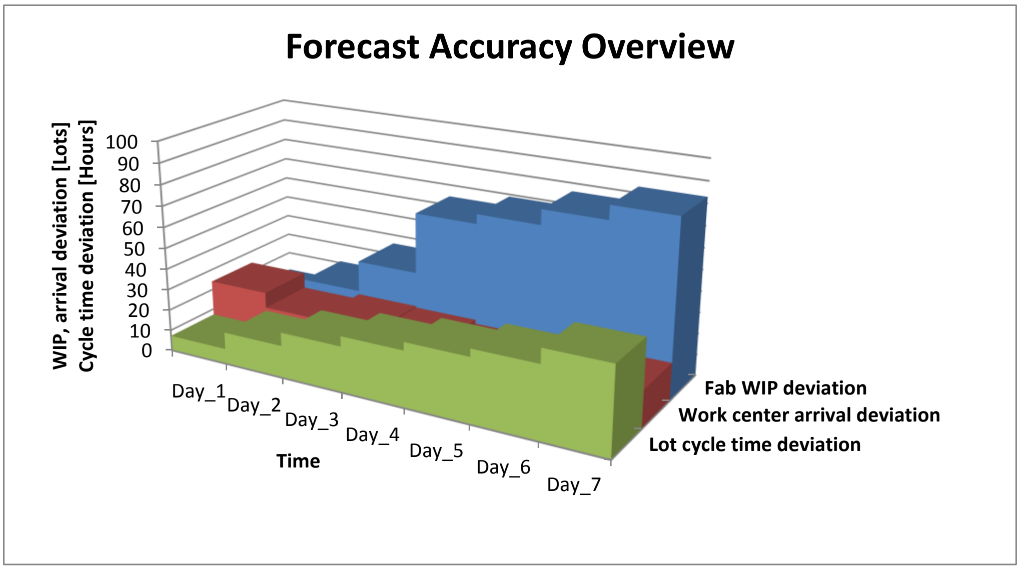


Figure 135: Forecast accuracy overview with absolute values

The conclusion is that a single chart, either with absolute values or relative values, is not suitable to represent the quality of the forecast results. For different levels of detail, the used KPI values and the computation formula are very different and not comparable. The scaling highly depends on particular value. The recommendation is to evaluate the forecast quality with different KPI for different levels of detail in separate charts.

Even if the KPI and their computation formulas are clearly defined in this thesis, the complexity of the generated results is very high. For a common use of a single chart, there is the danger of a totally different interpretation, even if the data values are the same. The interpretation highly depends on the particular parameter, its computation formula, its unit and the scaling of the charts.

6.6.3 Simulation Accuracy

The analysis of the online simulation accuracy shows the mean absolute deviation between the forecast and the simulation results. Multiple levels of detail have been analyzed.

Fab level

On fab level, the deviation (MAPE) of the fab WIP is around 2%. The deviation of the fab moves is around 11%. The accuracy on fab level is quite close to reality. The reason for the deviation of the fab moves is the overall model behavior, which is slightly too optimistic. In a mature high mix logic fab, many sources of disturbance exist. Many disturbances are not part of the simulation model. The online simulation model is also very stable over a long time period. Even for a forecast horizon of more than seven days, the fab moves and the fab WIP are very close to reality.

Work center level

On work center level, the accuracy deviation for WIP and arrivals has been analyzed. The deviation of the work center WIP is very high. It reaches a level of more than 100%. The WIP forecast is highly affected by stochastic effects but also by bad data quality. It is not recommended to use this forecast in the production environment as a standard report. Despite the results quality, the WIP forecast is useful for very few selected work centers. Especially for those work centers, where the stochastic influence is low, the data quality is good, and the lot arrivals reflect reality, the WIP forecast is still useful. For the work center arrivals, the results are much better. The deviation for seven days of forecasting for all work centers is about 40%. This number also contains many low volume work centers. For relevant work centers executing at least 1% percent of the fab moves, the arrival prediction is much more accurate. For these work centers the deviation is only 11%. So, a simulation based lot arrival forecast is highly useful to predict the future work center behavior.

Lot level

The lot level is the most detailed level of detail. An accurate forecast on lot level is far beyond any expectations. The results analysis shows that after a forecast horizon of 7 days, the absolute cycle time deviation already reaches a value of 1.7 days. The deviation is very high. Only 50% of the lots from reality find a corresponding matching partner in simulation to compare the cycle time. A more realistic time horizon for a lot based forecast is one day. The gap between simulation and reality on the first day, with a cycle time deviation of around 100%, is also quite high. Only 12 % of the lots in reality cannot be compared with the corresponding arrival events in simulation.

6.6.4 Stochastic Effects

The stochastic effects have been analyzed for different level of detail. For each analysis ten confidence runs have been executed. The results are as follows.

Fab level

On fab level the relevant KPIs are fab WIP and fab moves. The stochastic effects for the fab WIP are very low. For a seven day forecast horizon, the size of the confidence belt reaches a level of about 1%, compared to the average forecast value. For the fab moves, the stochastic effects are quite high. The size of the confidence interval reaches a level of about 7%. Another conclusion is that the size of the confidence belt is increasing over time. This is the expected behavior. The confidence interval is much smaller for the first couple of days until the maximum of 1% for the WIP and 7% for the moves has been reached.

Work center level

The stochastic influences on the work center WIP and the work center arrivals have been analyzed. For the work center WIP, the size of the confidence interval reaches a level of about 90%, compared to the average forecast value. WIP forecast is not usable due to large stochastic influence. For the work center lot arrivals, the influence is much lower. It reaches a level of 20% for all work centers. For online simulation in semiconductor manufacturing such a low influence contributes significantly to an accurate forecast result. After simulation start, the size of the confidence interval is also quite stable. Even on the last days of the forecast horizon, the size of the confidence interval is only slightly increasing.

Lot level

On lot level, the stochastic effects have been analyzed for the lot cycle time. For seven days of forecasting, the size of the confidence belt for the lot cycle time is about 1.2 days. It is caused by stochastic effects. This is an influence of about 17%. The absolute size of the confidence interval is increasing over time as expected. For a short time horizon, the stochastic influence is also tremendous. For a 6 hour lot arrival forecast, the size of the confidence belt is already 6 hours.

Conclusion

The conclusion is that the stochastic effects in the simulation model have a different impact on different levels of detail and different KPI. Especially for work center WIP, the influence is high. According to the expectations, the major reason for the stochastic effects is the equipment downtime behavior. To improve the forecast quality, a common way is to execute confidence runs and compute the average value. For the work center WIP, which is highly affected by stochastic influence, the expectation is that this strategy improves the forecast accuracy. The analysis of the stochastic behavior shows that this strategy upgrades the overall forecast quality by 6% only. An average result from confidence runs is only slightly more accurate than a single run. Therefore the decision is to avoid time consuming confidence runs and execute a single simulation run. A minor reduction of the forecast accuracy seems to be acceptable.

6.6.5 Forecasting Accuracy in Literature

In semiconductor manufacturing only few literature sources are available which provide actual numbers for the achieved forecast accuracy.

Bagchi et al. (2008) present various simulation results like tool throughput, fab throughput, WIP prediction per process type, throughput prediction per process type, and lot trajectories. Multiple figures and examples are available to present the forecast accuracy. The only published value regarding the forecast error is the deviation of the weekly X-factor (flow factor). The achieved accuracy to forecast this value is less than 5%.

Zisgen et al. (2008) uses a fluid model approach. The main purpose is capacity planning and forecasting. Therefore the WIP and the X-Factor are in the range of interest. This approach is reliable to predict the fab WIP and the fab X-factor with a forecast error of less than 10%.

The analytical method from Mosinski et al. (2011) provides a lot arrival forecast for an operation in the wafer fab. The achieved accuracy is higher than 90% for the first three days and still above 88% for the first week.

These literature sources provide a short overview of the achieved accuracy for performance forecasting in semiconductor manufacturing. The accuracy highly depends on the particular fab environment, the forecast parameter, and the level of detail. A comparison of the forecast methods itself and the related accuracy is not useful due to such heterogeneous conditions.

For future research the recommendation is to define a common standard for the forecast results. A comparison of forecast methods and a comparison of different fab environments is only possible, if the generated results have the same level of detail, the same KPI, and the same computation formula to measure the forecast error.

7 Applications

Operational control is highly important for efficient manufacturing. Operational decision makers change certain fab settings to meet the requirements of the production process. This is an ongoing process. A short term simulation forecast supports operational decisions with the objective to improve the overall fab performance. Thus the operational decisions are the underlying applications of online simulation. All of these operational applications define the requirements of online simulation. The first part of this chapter lists many examples for operational day-to-day decision making in semiconductor manufacturing. These examples are described in detail. In the second part, the system requirements are defined.

7.1 List of Applications for Operational Decision Support

For online simulation the following examples of operational applications exist. They have been identified during interviews with the productions departments but also in the literature (Kohn et al, 2009). It is obvious that many underlying applications exist. These applications affect different areas in the fab like the maintenance, the production control, and fab operations. The examples in this list are described in detail in the following sections:

- Preventive maintenance forecasting
- Preventive maintenance planning
- Execution of engineering actions
- Resource saving
- Reticle cleaning
- Backup tool activation
- WIP balancing
- Dispatch rule selection
- Simulation based scheduling
- Lot arrival forecasting
- Sampling changes
- Wafer out prediction
- Lot release
- Setup changes
- Bottleneck detection
- Dedication and qualification

It is necessary that the online simulation capabilities meet the application requirements. To measure compatibility of the simulation forecast and the operational decision the parameters “level of detail” and “time horizon” are used. Regarding the level of detail some applications need results with a very fine granularity, while other applications need abstract results. A different level of detail for model resources is for example the fab level, the work center level, the equipment level, and the chamber level. Time precision is also an issue with regards to the levels of detail. It is necessary to define if the effective time unit is seconds, minutes, hours, or days. Another parameter is the time horizon. Each operational application has a certain time horizon. The application decision time horizon is the duration between the point in time where the decision is made and the time when this decision takes effect. This decision time horizon needs to be smaller or similar to the simulation forecast horizon, so that the forecast has full effect.

7.2 Description of Application Examples

Preventive Maintenance Forecasting

The first interesting model application from this list is the PM forecasting. In semiconductor manufacturing basically three kinds of PMs exist:

- Fixed PM period
- Lot counter based PMs
- Process parameter based PM

First, for equipment with a fixed PM period, the time between two PM's is always the same. A PM action based on a lot counter means, that after a certain amount of lots the equipment stops processing. The equipment enters a scheduled down state automatically. Equipment PMs based on process parameters stop processing if the process parameter is out of its specification. In practice the process parameter trend correlates with the number of processed lots. This characteristic is quite close to counter based PMs. An example is the particle density parameter. This value increases for subsequent plasma processes in an equipment chamber. This PM type is an intermediate type between scheduled an unscheduled down. The PM forecasting becomes interesting for PM types based on lot count. Without a proper lot arrival forecast, the point in time when the PM occurs is unpredictable. A PM can take place three days earlier during high loading conditions for example. During low loading conditions it can happen three days later. Therefore it is hard to assign the maintenance staff and to order the material at the right time. In some cases a PM occurs too early, when the spare parts and the maintenance staff are required but are not available yet. In few cases a PM is delayed and the assigned maintenance resources are not yet required. This reduces the fab performance. The objective of a PM forecast is to increase planning security for a better fab performance, and to save resources. To do so, the requirement is to have a lot based forecast on equipment level or even chamber level. The desired time horizon is a few days up to two weeks. The scale unit for time is one shift because the maintenance department is organized in shifts. In research, the topic of predictive maintenance is highly up to date (Dietmair et al. 2010). A lot arrival forecast contributes to achieve better results in that area.

Preventive Maintenance Planning

Another application is the preventive maintenance planning. This topic is closely related to PM forecasting. PM planning includes all types of PMs mentioned above. The challenge for this application is to increase the work center performance by planning upcoming PMs according to future workload. The problem is that in complex manufacturing environments, a workload difference occurs very often because of the high variance in the wafer fab (Hopp and Spearman 2000). Often a PM is planned independently from the future workload. The result is that the maintenance department executes a PM even if the work center workload is high and a different better point in time would be available to execute this PM. The result is that the PM increases the work center cycle time and the work center WIP. The improvement is to execute the PM when work center WIP and work center lot arrival is low. It leads to a reduction of lot cycle time and work center WIP. The PM planning horizon varies from 1 day up to 1 month. The required levels of detail are the work center WIP and work center arrival. This topic also has a lot of potential for what-if scenario and optimization. The control parameter is the time of the PM. The objective is to reduce the WIP and the cycle time of that work center.

Engineering
Beside the normal production the equipment is used to process special tasks. The equipment processes non-productive lots. The term “Engineering” is used to define the status of this kind of tasks for the equipment. Examples from the manufacturing area are the following:

- Development of new processes
- Development of new products
- Equipment qualification

Some of those engineering tasks require long time. It is feasible to plan those actions according to future work center workload to avoid delays for productive lots.

Resource Saving
Another use case is resource saving and energy saving (Dietmair and Verl 2010). The key challenge is to switch off equipment or equipment components without a reduction of the fab performance. The time to shut down and to reactivate these components becomes important. Multiple shut down stages are available. The reactivation of several electronic components only takes a few seconds. It needs several minutes to switch a computer on and off. A heater station or cooling station takes even longer. To reactivate a particular process on the equipment, it needs up to several months for reactivation, qualification, and testing. Another aspect is the change intervals of spare parts, wearing parts, and consumables. A higher change interval leads to resource saving.

Multiple options are applicable to accomplish resource savings. First option is to apply lot scheduling. Due to a high uncertainty in a wafer fab this option is available only for a very short time horizon, such as a few hours. Another option is to predict the future workload to decide for how long it is feasible to switch off the equipment. With the knowledge of future workload it is feasible to affect the change intervals of spare parts. The third option is to consider shut down states in long term capacity planning. Other operational problems in a complex wafer fab, also deal with energy saving aspects, such as equipment dedication and lot scheduling.

The requirements highly depend on the desired resource saving approach. For energy saving on operational level, the desired level of detail is equipment level or equipment component level. The overall future workload also becomes important. The time horizon and the time precision highly depend on the time needed to switch on and off a component. This takes a few seconds up to several months. Detailed resources saving examples are the following:

- Switch equipment components on and off
- Interval of bath changes for wet benches
- Interval of bulb changes for lithography
- Dedication at lithography whereby each process needs a different coating

For the first two applications it is clear that the resource consumption depends on the length of the time interval to replace the components. A forecast helps to determine the point in time, when for example the replacement for something on a lot counter is necessary. This application is closely linked to the PM prediction topic. The last example refers to the number of dedicated equipment. For example to change the dedication, a new bottle of fluid replaces the old one with a different type of fluid. The waste of resources is large because a large amount of fluid flows through the filter, to reach stable process conditions. From the resource savings perspective it is best to limit the amount of dedicated equipment and limit the resource requirements. From the fab performance perspective it is best if no dedication exists and the work center equipment is able to run similar processes. Simulation helps to determine the tradeoff between the minimum number of dedicated equipment and the minimum performance loss.

Reticle Cleaning

Another application is to plan the time for the reticle cleaning. From time to time a reticle needs a cleaning action. For this time, the reticle is not available for production. The best time for such a cleaning is when the workload for a particular layer and for a particular product is low. The target is to clean the reticle without cycle time loss for the lots. A simulation based forecast helps to figure out the best time for such a cleaning.

Backup Tool Activation

The backup tool activation scenario is also an application with resource saving background. Due to the high importance of this application, it is worth to mention it in a separate paragraph. Without a forecast, backup tools are activated, if the WIP for a work center becomes too high.

The disadvantage of this approach is that a lot of WIP has to build up before the backup tool is used. In this time the WIP increases even further because of the delay time between the WIP increase and the activation of the backup equipment. A simulation based performance prediction helps to activate backup tools early to avoid a large WIP built-up.

WIP Balancing

Another application for online simulation is a dynamic WIP balancing approach (Sivakumar et al. 2008). The background is to reduce the variability by smoothing the WIP situation in the wafer fab. With online simulation it is feasible to configure the WIP load level and the release control points in the fab.

Dispatching Rule Selection

A dynamic dispatching rule selection and configuration is applicable in environments with high variability. Wu and Wysk (1989) present a simulation based dynamic dispatch rule selection method. For a short time period, a variety of dispatching rules has been simulated. For the following time period, the wafer fab applies that rule with best performance in simulation. To activate this application, it is necessary to create a simulation model and the simulation based optimization feature. It is necessary to define, which dispatch rules and dispatch rule parameters are available for optimization.

Simulation Based Scheduling

In addition to dispatching, simulation based scheduling solutions are also available. Horn (2006) and Potoradi et al. (2002) apply this approach in the semiconductor backend and in an assembly line. To consider lot scheduling solutions, batch creation, preventive maintenance scheduling, and setup scheduling are also a part of the scheduling problem. Due to the high variance in a front end wafer fab, this solution is applicable for a very short time horizon. The timing scale unit is minutes. The required level of detail is lot level and equipment level. The simulation based forecast is the first step to enable simulation based scheduling. The online simulation system needs to be extendable for what-if scenarios and optimization.

Lot Arrival Forecast

Besides simulation based scheduling, other scheduling methods are available (Klemmt et al. 2008). Many of those mathematical work center optimization methods require starting values for example the estimated arrival time of a lot. Online simulation is capable to provide a lot arrival forecast. Due to high variability in a wafer fab, the maximum time horizon for scheduling solutions is very short, ranging from a few hours up to one shift. Time scale unit for the forecast is in minutes. For each lot information about the executed operation and the assigned work center are also required. In industry and science, the interaction of simulation and optimization solutions is in use. An example is that simulation provides the starting values for optimization (März 2011).

Sampling

Sampling optimization is an application for wafer testing in a front end wafer fab. Sampling defines, if a lot executes an optional measurement step or not. The sampling rates are flexible. Low volume products, especially new products with critical processes need much more testing than high volume products, where the processes are quite established. If many lots of the same type arrive at the same time, not every lot needs a measurement. Only a few lots are required to collect the desired measurement values. So it is feasible to connect the concept of volume dependent sampling with online simulation. A lot arrival forecast helps to adjust the sampling rates early. The reason behind this is that sampling rates become important if high WIP turbulences affect measurement operation. A measurement work center becomes a bottleneck in the fab, even if this measurement work center is not critical at all according to static capacity planning. The required level of detail is the product level per work center. The time horizon ranges from a few days up to one week.

Wafer-Out Forecast

Another application is the wafer out forecast. According to Robinson (1998) the semiconductor manufacturing supply chain consists of four basic steps: wafer fabrication, wafer test, assembly, and final test (Figure 136). A wafer out forecast becomes important for several operational decisions in the downstream facilities of the supply chain. To predict the upcoming loading of the wafer test facility, the wafer out forecast of the front end wafer fab is necessary. The required level of detail is product level. Time horizon ranges between several days up to multiple weeks.

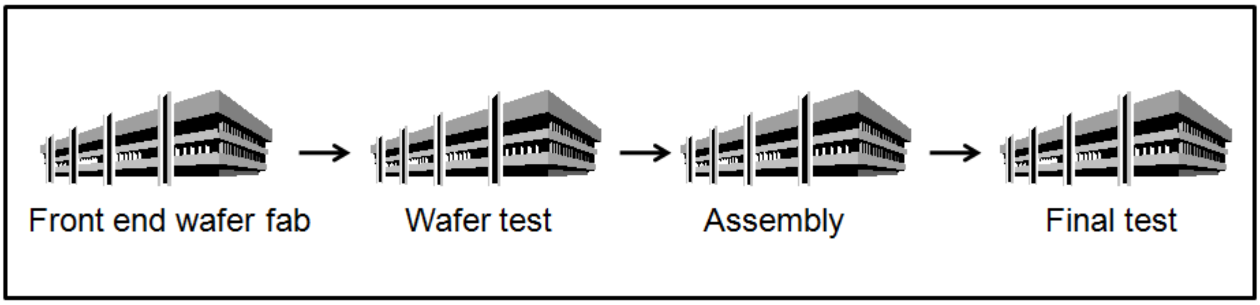


Figure 136: Manufacturing facilities in semiconductor supply chain

Lot Release

One application is meant to improve the lot release on operational level. The objective is to release a given lot volume over a period of time in an optimal way according to the fab performance. As seen in Figure 137 and Figure 138, the current lot release from one source of one week is not equally distributed over time. Most lot releases occur between 8 and 11 am. The lot release volume also increases from Saturday to Thursday. This kind of unbalanced lot release leads to disturbances from the beginning onwards. An intelligent lot release strategy balances these disturbances early. Consideration of the batch size and the WIP lots of the work center at the start of the step sequence leads to further improvements.

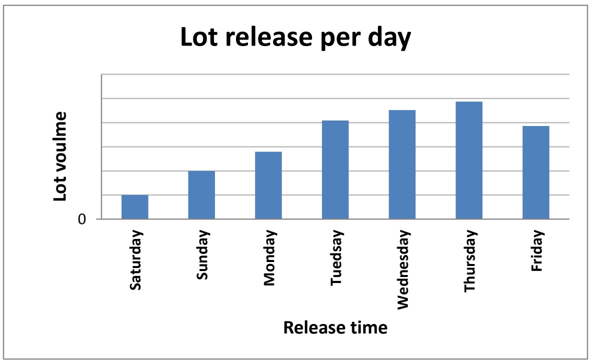


Figure 137: Avg. lot release per day

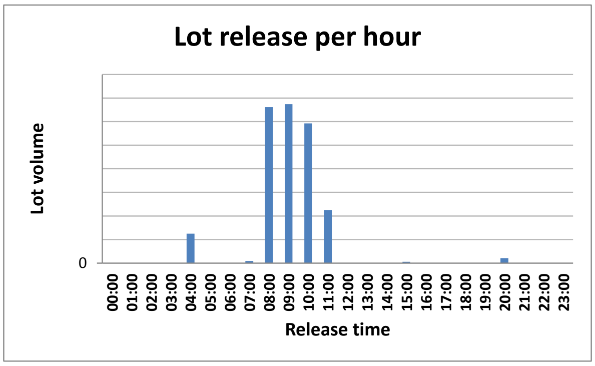


Figure 138: Avg. lot release per hour

Setup

Setup decisions are important for those work centers where setup have a high impact on equipment throughput. The decision to change the setup of the equipment is very complex. A common reactive method is to wait until a certain amount of WIP is waiting to execute the setup change. The disadvantage is that much WIP needs to cumulate and wait until setup is finished. The result is a cycle time loss. Regular setup times of only a few minutes are not in the range of interest. Setup times, which take several hours, have large potential for performance improvement.

A lot arrival forecast and WIP forecast are required on process level. The maximum forecast horizon is two days. An example for a setup decision is the implantation area. Dopant changes take a very long time. The throughput of implant equipment is also limited, because it is a single wafer tool.

Bottleneck Detection

Another application is the bottleneck detection. The background is that the bottleneck limits the throughput for the whole system. In a dynamic wafer fab environment with a high product mix, multiple bottlenecks exist and the bottlenecks are often changing. Still some equipment is more prone to affect the throughput for the system than other equipment. According to Roser et al. (2001), it is necessary to find the bottlenecks first before it is possible to improve the throughput of the bottleneck. Therefore it is highly useful to detect the operational bottlenecks. It is feasible to assign additional resources to the bottleneck equipment. Examples for additional resources are more operators to reduce the equipment idle time, or more maintenance personell to reduce the equipment downtime duration.

Dedication and Qualification

It is also achievable to connect dedication and qualification with online simulation. Depending on the future arrival situation, it is possible to optimize the equipment dedication of a work center. The objective is to reduce the queuing time of the lots which is caused by dedication constraints. Another objective is resource saving which has been mentioned before.

The feasibility of the dedication application depends on the decision horizon. The decision horizon for dedication and qualification varies very much from less than one week up to several months. The time duration depends on the particular process. To handle this application, the time horizon of online simulation needs to be larger than the time which is necessary to change the dedication. Online simulation additionally needs to consider dedication and it needs to be extendable to optimization. An arrival forecast which is sensitive to the process decision is necessary.

Summary

This thesis addresses several applications for operational decisions making. There is still the expectation that multiple other applications exist. Only system experts like operators or foremen with deep domain knowledge are able to define additional applications. So it is feasible that the generated results of the forecast system are used in many other situations on the shop floor.

7.3 Identification of Operational Applications

The next step is to identify a few applications which are implementable in the wafer fab. Therefore the list of applications from the previous section is a good starting point. In addition it becomes necessary to specify those applications. The following section provides an overview of the criteria to specify the applications. The second part describes the process to identify operation applications. The selected applications with their specifications will also be presented.

7.3.1 Application Specific Reporting

There are many criteria to specify the forecast requirements of an application. Those criteria are tightly connected with the reporting specification of the forecast results. The reason is that the reporting specification is the interface between the simulation results and the application requirements. Therefore this section discusses the reporting elements to find useful criteria to specify and to select the feasible applications.

Multiple literature sources provide an overview on reporting of wafer fab performance measurement. Hopp and Spearmen (2000) use the metaphor of a factory cockpit to describe that multiple performance indicators are required to reflect the fab performance. Standards are established to create common productivity measurement variables (Semi 2003 a). For each parameter, a clear description, the relations between all parameters, and the computation formulas are available. From simulation forecast point of view, several parameters are highly useful. Thus, this section discusses the relevant forecast parameters and the reporting criteria.

Each report consists of several components. This section contains an overview of the useful elements to create an appropriate reporting for online simulation. Following reporting elements exist, like the particular parameters, the categories, the aggregation function and the diagram type (Table 40).

Parameter value • WIP • CT • Moves	Categories • Resource object • Products • Time unit
Aggregation function • Average • Sum • Median • Count • Min • Max	Diagram options • Histogram • Pareto chart • Bar charts • Pie chart • Line chart • Table structure • Color indication

Table 40: Reporting elements

In a real wafer fab, the reporting categories are interesting because many categories and many elements within these categories exist. Table 40 contains a list of the reporting elements and several examples. Within the categories, many elements exist (Table 41). To express trend changes in a report diagram, it is necessary to select the appropriate time unit category. It is applicable to report a single parameter per hour, per shift, or per day. For the resource category it is also useful to generate reports for the full fab, for a work center, or for a single piece of equipment. Beside those grouping criteria mentioned in Table 41 many other criteria exist in practice. Examples are the number of wafers per lot, the route of the lot, the lot diameter, or the rework state.

Time Unit Category	Resource Object	Products
• Week • Day • Shift • Hour	• Full fab • Department • Work center • Equipment • Chamber	• All Products • Product groups • Single Product

Table 41: Reporting categories

Even simple categories like products, are very complex, if the product category consists of a large hierarchy tree (Figure 139). On top level, about 10 different main product groups exist, while on the lowest level more than 1000 different products exist. Therefore it is necessary to determine useful reporting categories for user convenience. If the category is too detailed, the report contains too many elements, such as 1000 product values. The result is not end-user friendly due to too much information. If the category is too abstract, the information content is very low and the report becomes worthless.

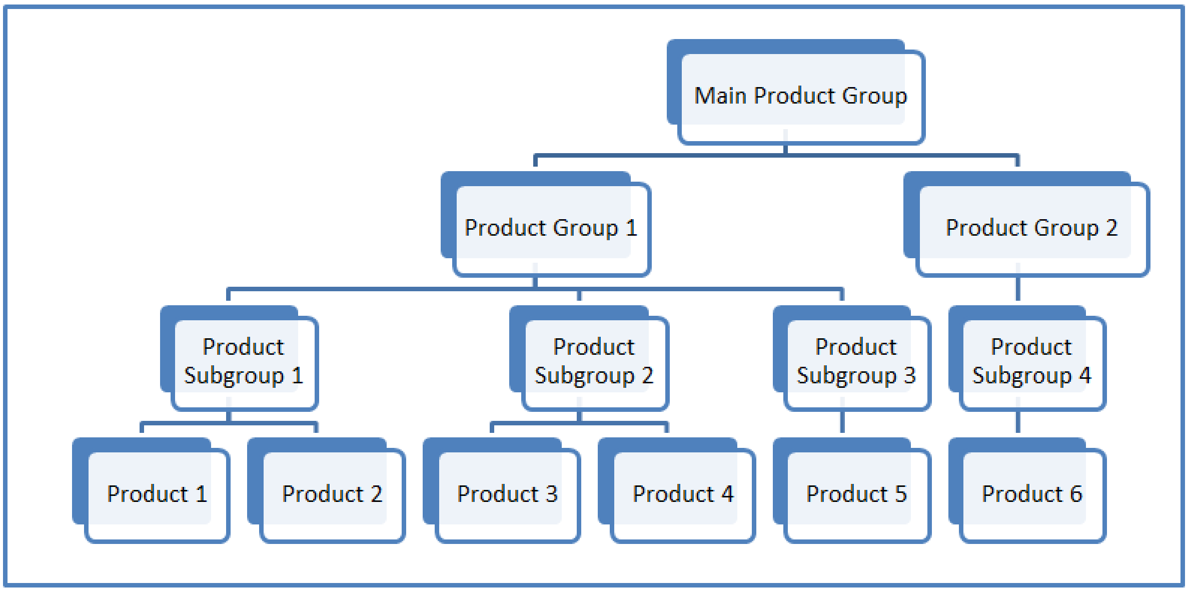


Figure 139: Product hierarchy tree

For each category it is necessary to apply powerful filter strategies. For example for the time period it is useful to display only the values within the forecast horizon. For the resource category it is useful to display only equipment information from a single work center and filter out all equipment from other work centers. For complex hierarchical categories it is also useful to apply filters for the right level of detail.

7.3.2 Identification of Operational Applications in the Wafer Fab

Based on the reporting criteria from above, a questionnaire has been developed to identify feasible operational applications. Those applications have specific requirements. The criteria to define their needs are the following:

- Forecast parameter
- Time horizon
- Time granularity
- Grouping criteria

Overview Requirements

Several department interviews have been conducted to identify the application requirements. For each department one pilot work center has been selected. The purpose is to figure out feasible applications and specific requirements. Examples for these requirements are the forecast parameters themselves, the required forecast time horizon, the grouping criterion, and so on. An example of a table based question form and the related results is available in Table 42.

Issue	Work center I	Work center II	Work center III	Work center IV	Maintenance	Industrial engineering
Forecast parameter	WIP, Arrival	WIP, Arrival	WIP, Arrival Product waves with Ranges	WIP, Arrival	Number of processed wafers	Lot arrival
Time horizon	1 week	1-2 days	1 week	1 week	1 week, minimum of 2-3 days	Process time
Time granularity	Per day	Per day	Per day	Per day	Per day	Per minute
Grouping criteria	Process	Process	Product group	Differentiate between the two main group of processes	Equipment, chamber	Work center
Forecast enhancement	Attribute modeling	Test wafer modeling	Consideration of high process times for lots with high priority	Kanban and test wafer modeling		Detailed initialization
Application	PM planning, PM prioritization, Special measurement	PM planning	Backup tool activation, planning of engineering actions	Backup tool activation, dedication changes	PM planning	Lot scheduling, dispatch rule config.

Table 42: Reporting requirements table from department interview

Regarding the requirements it can be seen that in the column for the maintenance department, the number of processed wafers is in the range of interest. The desired categories are the equipment level and the chamber level. The time granularity is one day. So for each day within the forecast horizon, one value is available. The forecast horizon needs to be one week. The minimum forecast horizon is at least 2 or 3 days, to make use of the simulation results. The requirements for the other work centers are different. For example for work center III, the WIP and the lot arrival are in the range of interest. It is necessary to report the amount of products in transit between two operations. The grouping criterion is the product group, where one single product belongs to exactly one product group. The particular application is to support the backup tool activation and the planning of engineering actions. Simulation forecast improvements are approachable through consideration of large process times of highly prioritized lots.

Besides the work center level, custom requirements also exist for applications like PM planning and lot scheduling. It is necessary to specify the requirements and the related interface definition.

Requirements PM Forecasting
The forecast helps to predict counter based equipment PMs for the maintenance department. The number of processed lots determines the point in time, when the PM is necessary. Several options exist to meet the needs of PM forecasting:

- Number of lots per equipment
- Equipment processing hours
- Time of the next equipment PM

Instead of reporting the number of processed wafers, it is also feasible to compute the processing hours of the equipment. Another alternative is to include that information into the simulation model and integrate the preventive maintenance forecast directly. All these options are feasible, depending on the effort and the benefit.

Requirements Lot Scheduling
For lot scheduling the requirement is to provide an arrival forecast on lot level. This forecast is used as an input parameter for lot scheduling. Therefore it is necessary to provide the arrival time of all lots at a particular operation for a short time horizon. This particular simulation output is an interface definition to subsequent applications like lot scheduling. Table 43 shows an example of this type of output.

Lot	Operation	Enter Operation Time Stamp
Lot_01	580	29/08/2011 05:26:41
Lot_04	130	29/08/2011 06:21:12
Lot_01	590	30/08/2011 07:46:41

Table 43: Lot arrival forecast

It is evident from the examples above that each application has specific needs. The simulation forecast and the related reporting have to meet these needs. The interview with the production departments shows that the major objective of short term simulation is the lot arrival next to the WIP prediction. Therefore, the paramount task is to predict trend changes, especially for lot arrival peaks. The production department is highly interested in the point of time where lot arrivals peak, the impact of the peak, and the length of such a peak.

In the department interviews one important reporting fact is missing. During the interview it is not achievable to define the acceptable degree of forecast error. The question is how accurate the forecast results have to be to become or remain useful. This is not part of the interview because the required deviation as well as the efficiency is hard to quantify. Every user demands perfect forecast results. The feedback from the practitioners is that the forecast results are good enough, if they are used by the production departments. Within the scope of this thesis, the next section discusses the relation between the forecast error and the advantage of such a forecast for particular applications.

7.4 Accuracy of Results for Online Simulation Applications

To show the benefits of the simulation forecast, it is useful to compare the forecast properties with the requirements of each selected application. The previous section identifies the lot arrival forecast application, the PM planning application, the backup tool activation application, the dedication application, and the dispatch rule selection application. To compare the requirements, the criteria are: the computation time, the forecast horizon, the level of detail, and the forecast error. The following description indicates which application is feasible and which is not.

Lot Arrival Forecast for Scheduling

For a lot arrival forecast the created simulation solution is not applicable. The provided level of detail and the available forecast horizon is compatible with simulation results. The problem is the large computation time and the large forecast error. The computation time takes slightly more than 15 minutes, whereby the requirement is to obtain a solution within the process time of the equipment. The typical process times range from 15 minutes to 8 hours. A vast proportion of the forecast horizon is wasted due to the computation time. Regarding the forecast error, the cycle time deviation of the first day is enormous. The influence of stochastic effects on lot level is also huge. Within the first 6 hours, the size of the confidence belt already reaches 6 hours where the mean absolute forecast error reaches 3 hours. The magnitude of error is too large for a reliable forecast.

PM Planning

For PM planning the simulation forecast is applicable. The level of detail and the computation time match the requirements. Most PMs match with the forecast horizon. Only very few PMs have a long planning horizon. For those long PMs with external companies involved, a forecast horizon of 7 days is not sufficient. The achieved accuracy fits the needs of PM planning. The forecast error for work center arrivals of less than 11 % is a good result.

Backup tool activation

For backup tool activation, the described arrival accuracy meets the needs for this application. The level of detail and the computation times are also sufficient. The only constraint is the time horizon. A forecast horizon of 7 days is not suitable for those pieces of equipment where the activation time is longer than this.

Dedication

For the dedication application, the same constraints apply as for the backup tool activation scenario. The positive effect of a forecast will be reduced if the time to change the dedication is longer than the duration of forecast horizon.

Dispatch Rule Selection

For the dispatch rule selection the criteria are also in agreement with the specific needs, especially in terms of the achieved accuracy. The lot arrival forecast is capable of affecting the dispatch rule configuration for a time horizon between 1 day and 7 days. The specific interaction of simulation and dispatching depends on the particular implementation of the dispatching rule. The following section presents an example scenario for the dispatch rule configuration. The overview in Table 44 summarizes the feasibility of the applications. The table depicts the four criteria to derive the feasibility of an application.

Application	Forecast Value	Feasible	Computation time	Forecast horizon	Level of detail	Forecast error
Lot arrival for Scheduling	Lot arrival	No	No	Yes	Yes	No
PM planning	Work center arrival	Yes	Yes	Yes	Yes	Yes
Backup tool activation	Work center arrival	Yes	Yes	Yes	Yes	Yes
Dedication	Work center arrival	Yes	Yes	Yes	Yes	Yes
Dispatch rule selection	Work center arrival	Yes	Yes	Yes	Yes	Yes

Table 44: Overview of forecast applications and the feasibility

7.5 Example Scenario

This section presents an example scenario for the dispatch rule selection application. The purpose is to demonstrate how a simulation based forecast is capable of increasing the fab performance. To select such an example scenario, the following criteria apply. First, it is necessary to select a work center, where the overall work center performance is much lower than in regular work center behavior. The second criterion is that it is feasible to improve the work center performance by changing certain dispatching rule settings.

For the following example, both criteria apply. First of all, the work center performance is much lower than the average behavior (Figure 140). The planned WIP is around 1000 wafers. The WIP around day 10 exceed the planned WIP. It reaches a level of about 3500 wafers. Secondly it is feasible to improve the work center performance. In Figure 141 the throughput histogram of the real lot trace is depicted. A value of 120 occurrences at 11 min/lot represents 120 lots which reach a throughput of 11 min/ lot. Two peaks are visible in the figure, where the throughput of both peaks is 8 min/lot and 11 min/lot. This value depends on the setup change for equipment of this work center. If a process has been changed, the required setup time is reducing the lot throughput. Figure 141 shows that 120 lots execute a setup change and 80 lots do not execute a setup change. So the setup time influences the work center performance. There is a lot of potential for improvement.

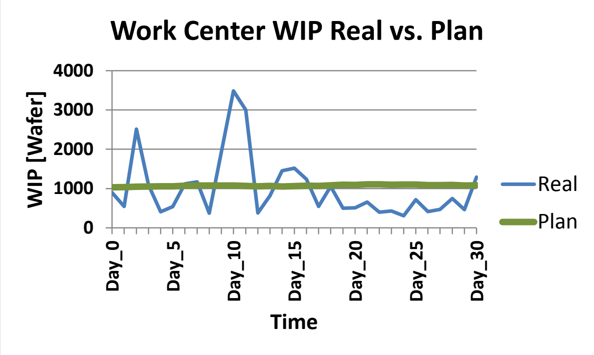


Figure 140: Work center performance

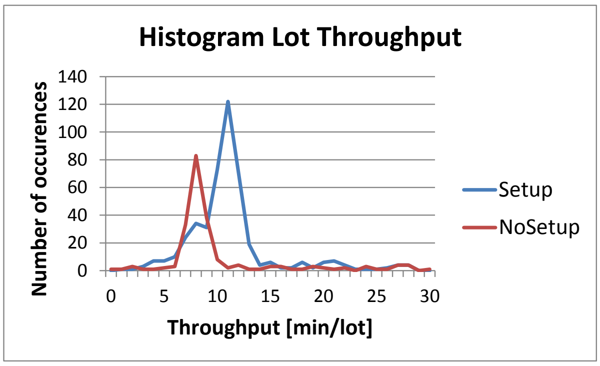
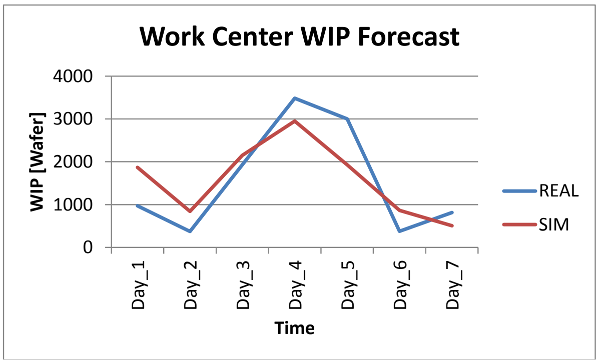


Figure 141: Histogram lot throughput

The simulation forecast results for this work center are available in Figure 142 and 143. The online simulation WIP and arrival forecast reflect reality. The time period for the forecast is the same time period between day 7 and day 13 from Figure 140 above. With the help of such a forecast, the production department is able to recognize problems early and react accordingly.



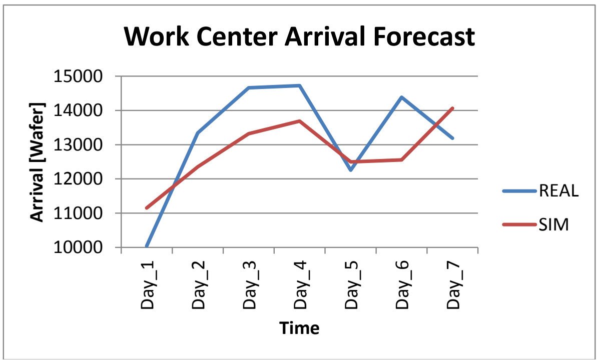
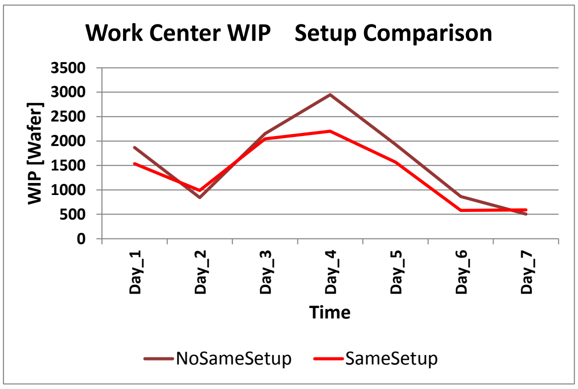


Figure 142: Work center WIP forecast

Figure 143: Work center arrival forecast

Knowing that setup has significant influence, the idea is to improve the work center performance by changing the same setup dispatching rule. The feedback from the production department is that this dispatching rule is applicable for this work center. The same setup rule has not yet been applied. Figure 142 and 143 show the work center WIP and the work center cycle. The figures depict two scenarios, without any changes and with the same setup rule changes.



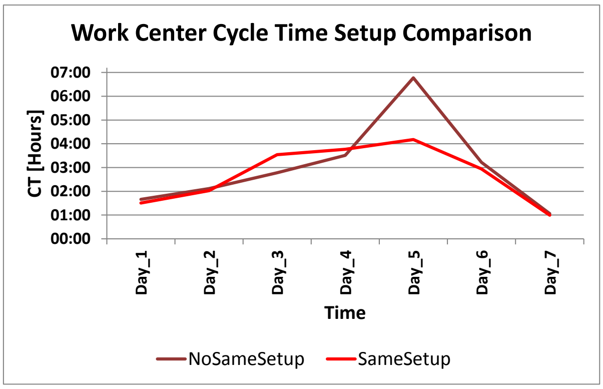


Figure 144: Work center WIP and setup

Figure 145: Work center cycle time and setup

In Figure 144 the work center WIP increases up to 2950 wafers in simulation without any dispatching rule changes. With the same setup rule, the WIP increase is much lower with about 2200 wafers. The same behavior is on display for the cycle time in Figure 145. Without the same setup dispatching rule, the maximum cycle time reaches a level of almost 6:46 hours. With the same setup dispatching rule the maximum cycle time is about 4:10 hours. To verify the results and to compare both scenarios, 10 confidence runs are executed.

The conclusion is that a short term simulation forecast is able to improve the work center performance, and so fab performance. The effect of the simulation forecast results in combination with applicable changes in the fab is large. The example shows that the maximum work center WIP has been reduced from 2948 wafer to 2200 wafer, which is an improvement of 25%. The peak cycle time has been reduced from 6:46 hours to 4:10 hours, which is an improvement of 38%. When online simulation will be applied to all work centers in the fab, the expectation is that the improvement of the overall fab performance is significant.

8 Fab Driven Simulation

As mentioned in chapter 4, an acceleration of the data transfer and the model initialization is highly important to reach real time capability. A strategy to reduce the time for initialization of short term simulation is the fab driven simulation approach (Noack et al. 2010). The basic idea is to apply an ongoing update of the simulation model during simulation runtime. This method replaces the time consuming data transfer at the beginning of the simulation process. This chapter describes the prototype implementation of the fab driven simulation approach. It shows that is it feasible to start a well initialized simulation model instantaneously.

8.1 Intro

To make use of short-term simulation on an operational level, three aspects are essential. First, the simulation model needs to have a high level of detail to represent a small part of the wafer fab with sufficient precision. Secondly, the simulation model needs to be initialized very well with the current fab state. And thirdly, the simulation results need to be available very fast, almost in real time. Unfortunately these requirements contradict each other. It takes a long time to initialize a high precision full fab simulation model because of the massive amount of data. To overcome these time consuming limitations, the idea of the fab driven simulation approach was born. This chapter describes how to start a short-term simulation from the current fab state immediately, i.e. without further delay.

8.1.1 Problem Description

The characteristics of such an online simulation system compared to an offline simulation are as follows (Scholl et al. 2010):

- A short-term simulation run will cover a small time period (e.g. one week).
- The simulation model needs to have real time capability, i.e. to provide results before they are obsolete.
- The simulation model needs to have a high level of detail.
- The model initialization is important because the warm-up period is in the range of interest and not the steady state.

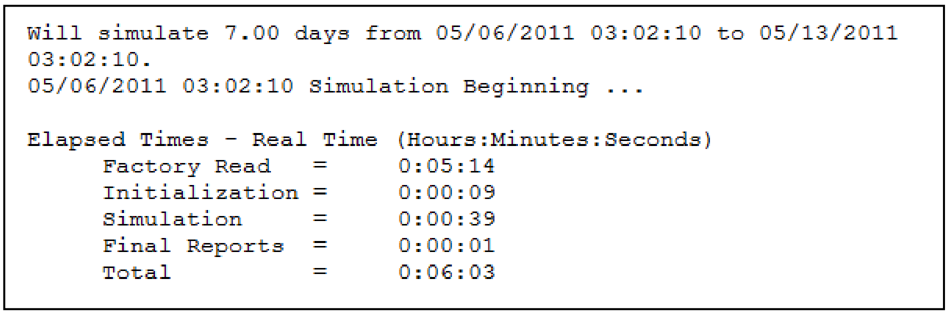
These properties have a large influence on each other, especially on the real time capability. As seen in Table 45, a very short simulation period is in agreement (+) with the real time capability. Hence the simulation execution time becomes very short. The high level of detail and the detailed model initialization is in contradiction (-) to the real time requirement. A high level of modeling details increases the simulation execution time. The importance of a detailed model initialization makes it necessary to extract a large amount of fab data at the beginning of the simulation.

Real time capability	+	Short simulation period
	-	High level of modeling details
	-	Importance of model initialization

Table 45: The effect of online simulation characteristics on the real time capability

These characteristics of an online simulation violate the requirements of a real time system. The time for data input and model initialization is very long, while the runtime of the simulation run is extremely short. The experience shows that the time for data input for online simulation takes much longer than the actual simulation run. Furthermore the major trend in semiconductor manufacturing is the increasing wafer fab size with an increasing number of equipment and lot numbers. This results in a rise of the data volume to initialize a simulation model. It has been shown that the data extraction, and model creation time is longer than the simulation execution time (Noack et al. 2010).

An example of a real short term simulation execution times is depicted in Figure 146. The time values differ between simulations runs depending on the particular configuration. In this case a single simulation run for a seven days period takes only 39 second, while the process to read the model takes 5:14 minutes.



```
Will simulate 7.00 days from 05/06/2011 03:02:10 to 05/13/2011
03:02:10.
05/06/2011 03:02:10 Simulation Beginning ...

Elapsed Times - Real Time (Hours:Minutes:Seconds)
     Factory Read   =     0:05:14
     Initialization =     0:00:09
     Simulation     =     0:00:39
     Final Reports  =     0:00:01
     Total          =     0:06:03
```

Figure 146: Execution time for a single ASAP simulation model run

So the fab driven simulation approach is highly useful for online simulation but not for steady state simulation. Only if the requirements of a highly detailed initialization and a short simulation horizon are available, the fab driven simulation approach has the biggest benefit.

8.1.2 Concurrent Solution Approaches

Besides the fab driven simulation approach, there are other methods to reduce the simulation model creation time. An initial approach is to differentiate the frequency of the fab data extraction. It is important to extract flow information, like lot moves, every time when a short-term simulation model is supposed to run. It is not useful to extract master data like route or equipment information with the same frequency. Therefore the update period is much shorter. A second approach applies to the model creation. Multithreading approaches make it possible to read simulation model information collaterally. At the moment most simulation engines read the model information like equipment, route, and lot information sequentially one by one. Both methods are applicable for online simulation as well. The disadvantage of these methods is that they minimize the data extraction and model initialization time, but they do not accelerate the model initialization time to instant availability.

Related work in this area was done by Aydt et al. (2008) in his studies on symbiotic simulation. A simulation of a wet bench work center is updated concurrently. The model update is done by a work center snapshot, where another simulation model emulates the behavior of a real wafer fab. This approach is very interesting because it is linked to this work in many ways. The major difference between Aydt et al. (2008) and the fab driven model approach is the update mechanism. While Aydt et al. (2008) use a full fab snapshot to update their model, the approach described in this thesis, improves the update performance by using single events. This approach reduces the amount of the update data volume significantly. This results in an improved scalability for the event based approach, which even allows updating a full fab model without violating the real time requirements.

Low et al. (2005) present an agent -based approach for managing all aspects of symbiotic simulation. The interaction of monitoring and controlling a physical system, the simulation and optimization and the whole system management is described in detail. This work is therefore interesting because the application domain is the semiconductor backend. Aydt et al. (2009) and Aydt (2011) summarize the work in the area of symbiotic simulation. A top down approach is used to investigate all aspects in that research area. Furthermore symbiotic simulation is applied in different domains like manufacturing and radiation detection. For this thesis the objective is to achieve real time capability for online simulation in semiconductor manufacturing.

Hanisch et al. (2005) provide an excellent conceptual overview how to synchronize a real world system with a simulation model. They distinguish several ways of initializing a simulation model. They present the conventional approach of model creation and initialization before the simulation run starts. The permanent model synchronization and "requested" driven model synchronization is also presented in detail. In a case study, they use the "requested" synchronization approach with the help of a simple train station model.

The approach of connecting external control logic with a simulation model also provides real time performance prediction and decision support capabilities (Smith et al. 1994, Mönch et al. 2003). The approach is quite interesting because the model synchronization and the real time performance prediction capability is a byproduct of the main intention. The main intention is to reuse the control logic (Smith et al. 1994). It is also a performance evaluation of shop floor control systems (Mönch et al. 2003). The control logic is an external module and it is not a part of the simulation engine itself.

Skoogh et al. (2010) present the first step towards the reduction of computation time. The idea behind is to continuously gather online fab data, over a long time period. This data is used as a simulation input. By updating this data continuously a significant reduction of computation time is achieved. The underlying fab data is always up to date and available at any time. So from online simulation perspective, the only time consuming steps are the model initialization and the simulation run. The fab driven simulation approach goes one step further and updates not only the underlying model data but also the simulation model during runtime. The time reduction does not only affect the data gathering process but also the model initialization time. A similar approach regarding the reduced computation time for the data input is available (Horn 2008). He uses caching and replication strategies to provide the required data in real time. Instead of capturing the data for the full fab state, the fab state from an earlier point of time is used. In addition an update captures the state changes since this previous point of time.

8.1.3 Motivation

The motivation to work on fab driven simulation is that the application of discrete event simulation is highly useful for operational decisions. The obstacle of a long initialization time is solved by using innovative methods. Therefore the fab driven simulation model has been developed. The reduction of model initialization time offers huge benefits for the model's application:

- Forecast results from simulation are not obsolete before they are available
- Immediate response to the user enables close user interaction
- Additional application areas for a very short time horizon become feasible (e.g. lot arrival prediction for lot scheduling)
- Higher initialization quality because of additional short-term initialization data (e.g. position of lots in transport system)

8.2 Concept

This section describes the basic principle of fab driven simulation. It shows how fab driven simulation changes the way to start short-term simulation runs in the future.

The modeling process of a regular online simulation run, without the fab driven simulation approach, is depicted in Figure 147 on the left side. The user needs to extract the data (1) and initialize the model (2) before starting a simulation run (3). This is necessary to obtain the latest fab state. The short-term simulation extracts the fab data automatically from distributed data bases. Several very time consuming data transformation, error checking, and error correction steps have to be executed. Furthermore the model initialization takes more time than the simulation run itself, depending on the particular scenario. All model files are loaded from the hard disk into the memory step by step.

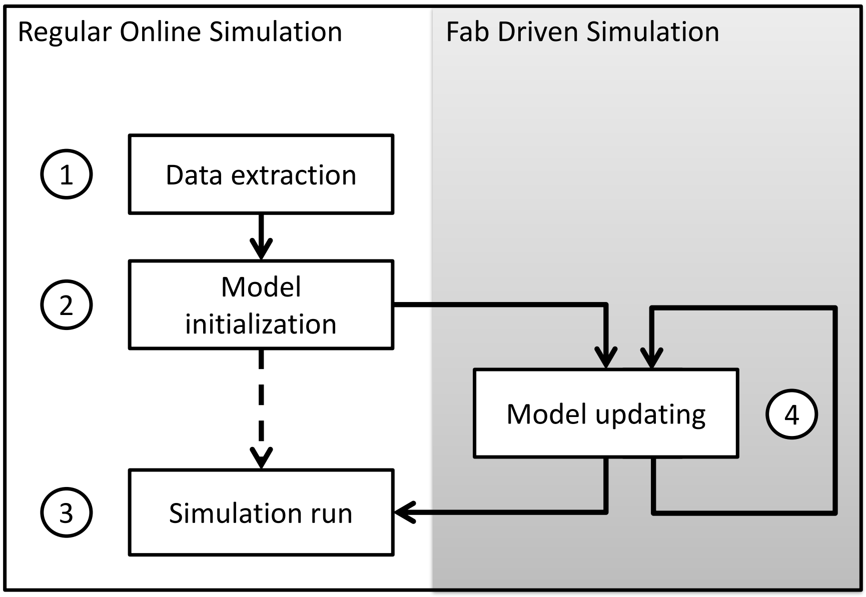


Figure 147: Process flow with fab driven simulation

With the fab driven simulation approach the model needs to be initialized (1 & 2) as well. But this initialization is done beforehand, for example overnight. After this point in time, the model will be updated continuously (4) on the right side in Figure 147. If the user wants to start the simulation (3), the most up to date model is already available. The user is able to run the simulation immediately and will obtain the results fast. Later on it is planned to duplicate the process instance of the fab driven simulation process before starting a simulation run. Thus the model update process will not terminate when the user starts to simulate. This approach provides a separate continuously ongoing update process at all times even during simulation.

8.3 Updating Requirements

The previous sections described the process of updating the simulation model, but it has not been mentioned yet what exactly needs to be updated. To figure out which update feature makes sense, the following two questions are important:

- Which event is related to the scope of the simulation model?
- Which type of event is useful to be updated regarding the duration of fab synchronization and the event update frequency?

The first question is easy to answer. The level of detail of the simulation model and its containing features define which elements need to be updated. If the level of detail for simulation is lower than the level of detail for the updates, the simulation model cannot handle most update events. Most events from reality do not have a counterpart in simulation. The opposite behavior occurs if the simulation model accuracy is higher than the update event accuracy. Many detailed events are missing in simulation. They cannot be synchronized because the update events from reality are missing.

The second question is more interesting. To answer the question what event type is useful to be updated, it is necessary to consider two elements. What is the average time, the simulation will run concurrently to the fab? And what is the average time period until the next event occurs for one event type? For example many lot moves occur during that time period, but only few station certification changes occur at the same time period.

8.3.1 Analysis Event Update Frequency

To evaluate what event needs to be updated, Table 46 contains an overview of common event types in a real wafer fab. Each event type contains the time intervals at which typical events occur. As it can be seen, the lot movement event type occurs about every 2.4 seconds while a new route is created only every 12.8 days. Thus, these times are an indication which event types have an impact on the model update horizon and on the model accuracy. As a proof of concept the decision is to update the lot moves because they are highly relevant for the simulation results and they occur very often.

Event type	Mean time between events
Lot moves	2.4 sec
Hold	11.8 min
Split	8.4 min
Unscheduled down	6.6 min
Scheduled down	1.3 min
New product	19.7 hour
New route	12.8 day

Table 46: The mean time between events t is computed by $t = t/n$, whereby n is the number of events and t is the observed time period.

Besides those events mentioned in Table 2, many other elements exist, which change over time. It depends on the level of detail of the simulation model whether those elements need to be updated. Events like dispatch rule changes are hard to quantify. Other changes are done in batches and not one by one. An example is the change of the lot release plans which will be updated once a week for most lots. These characteristics have a significant effect on the event update for additional model elements.

- Process time changes
- Dispatch rule changes
- Dedication changes
- Lot release changes
- Reticle information like reticle position, number of available reticle
- Equipment added or removed
- Sampling rate changes

Using the fab driven simulation approach, many more event types can be added to the simulation model. If the model initialization times and the simulation runtimes are long, these events and the related results are already obsolete before the simulation results are available. But when simulation results are available in real time, those events considerably improve the quality of online simulation for very short time horizons:

- Detailed equipment events are the start/finish of the process, the pump/vent, the heating/cooling
- Chamber events for cluster tools like 1st wafer finished, 2nd wafer finished
- Current lot position in the transport system

8.4 Design Decisions

From the conceptual point of view, multiple options exist to implement fab driven simulation. Thus, several decisions are required to identify the best solution to update the model. This section discusses multiple approaches and their advantages and disadvantages.

8.4.1 Update Trigger Approaches

This topic is concerned with the start of the update process. Two options are feasible to trigger a single update:

- An event in the fab triggers the update
- A periodical time interval triggers the update

The decision is to trigger the update regularly, within a predefined time interval of a few seconds. The advantage is a more predictable update cycle. It is easier to implement because the internal clock triggers the update and not an external event. Another advantage is to avoid the problem that multiple events trigger several updates at the same time.

Related trigger approaches from the area of symbiotic simulation are available (Low 2005).
The idea is to run the simulation and the emulation of reality simultaneously and monitor several resulting KPI. The simulation model is updated on demand, if the simulation and results from the emulated reality diverge too much.

8.4.2 Update Period

This section discusses the time between two updates. It is related to the length of the time period and the events contained. Furthermore, the real time capability of the data access is an issue. Two update types will be distinguished:

- Short update period, synchronous time stamps, real time data access
- Long update period, asynchronous time stamps, no real time data access capability

For the first type, the update period is very small, about a few seconds only. In this case the time stamps are almost synchronous, which means the time stamp of the latest update is similar to the time stamps of the events contained. The data access requirement is to immediately deliver all fab events that occur at this point in time. The advantage is that the simulation model is always up to date.

For the second type, the update period is very long, about several minutes, or even one or more hours. The time stamp of the last update and the time stamp of the events contained are not similar. It is necessary to distinguish these time stamps to schedule all simulation events at the right time. The advantage of this approach lies in a delayed data access which does not need to meet the real time requirements. Furthermore the system and network utilization is lower due to a reduced update frequency compared to the first approach. The major disadvantage is the need to manage both update time stamps and event time stamps. Another disadvantage is the gap between the fab state and the model state. This gap is caused by a higher delay time for the model update.

For this thesis the first approach is used because the model is up to date without much delay. The key to success is to achieve real time capability for the data access. To achieve this, the Real Time Dispatcher (RTD) from Applied Materials is used. It is the dispatching system of the wafer fab.

8.4.3 Simulation Time Synchronization Approaches

Two options exist to synchronize the simulation time with the real time:

- An external time stamp from the latest fab update is used in simulation.
- The internal simulation clock time runs with the setting of 1:1. One second of simulated time takes one second in reality.

The drawback of the second approach is the unsynchronized simulation and fab time and their deviation after a while. Typical reasons are minimal differences in the clock accuracy, mismatches during the leap year, and changes due to daylight saving. Furthermore this approach provides a wrong impression to the user, in a case where the fab event updates fail. If the simulation clock advances, the user thinks the model is still up to date, even if the model state is no longer synchronized and outdated. The conclusion is to use the first approach. The external time stamp from the last update event is used as simulation time. The simulation time will be increased in a loop until the new update time stamp is reached. The underlying assumption is that the time stamp of the next update is larger than the time stamp from the last one. If this condition is not valid, the simulation time will not be changed.

8.4.4 Handling of Simulation Events in Fab Driven Mode

For fab driven simulation two different sources for a single event exist, the simulation event and the fab event. To keep the model up to date, the fab event definitely needs to be included into the simulation model. The question is what happens with events from the simulation model. The following approaches exist to handle simulation events:

- Remove simulation events and block resulting state changes
- Allow simulation events changes and correct state changes later

The “Remove simulation events and block resulting state changes”-strategy removes the simulation event when the related fab event occurs first. If the simulation event is triggered first, the effect of this event will be disabled. The advantage is that the simulation is always synchronous to the wafer fab. The disadvantage is that many events are disabled over time. The second strategy allows regular simulation event execution and corrects state changes later. When the fab event occurs first, the simulation event will be removed as mentioned in the first strategy. In several cases, the simulation event occurs, before the fab event is

available. In this case, the simulation event will be executed, in contrast to the first approach which does not allow simulation event execution. If a fab event occurs after the related simulation event, the real fab event corrects the state change caused by the simulation event.
The decision is to implement the first approach which does not allow state changes caused by simulation events. If a simulation event occurs, the effects of this event will be disabled. There are two major reasons for this decision. First, the model is always synchronized with the wafer fab. Secondly, it is not necessary to implement a rollback strategy for simulation events.

8.5 Implementation

The implementation of the fab driven simulation consists of several modules (Figure 148). The following modules and their triggers are described in detail:

- Event export module: This module exports the fab events into the update file. It also exports the current time stamp of the fab. The export module is triggered after a predefined time interval. This is done by the system of the wafer fab.
- Event import module: This module imports the incoming fab events from the update file into the simulation. It furthermore imports the latest update time stamp of the fab. The file import is triggered by the simulation itself. Due to the fact that the import and export are not synchronized, the frequency of the data import must be higher than the data export frequency. It is necessary to avoid any situation where the export function overwrites events in the update file, which are not yet synchronized with simulation.
- Lot reallocation module: Based on the incoming events, this module integrates the lot movement events into the simulation model. This module is executed after the file import. Following main tasks are necessary to update the lot movements.
    - Check if the lot already exists.
    - Remove the lot from its current equipment or queue, if the lot exists.
    - Create a new lot if the lot does not exist yet.
    - Insert the lot into the queue, equipment or sink.
- Time synchronization module: It controls the time of the simulation. The purpose is to avoid a deviation between the simulation time and the time of the wafer fab. For this implementation the time stamp given by the import module is used. The time synchronization will be executed after the lot reallocation, when all other model state changes are done.
    - The simulation clock advances if the simulation time is shorter than the latest update time stamp.
    - The simulation time is stopped when the latest update time stamp is reached.
- Mode switching module: This module is necessary to terminate the fab driven mode and switch to simulation mode. The user decides when to start the simulation run. Therefore simulation events will be enabled while fab driven updates will be disabled. Following changes occur:
    - Disable update file import.
    - Disable lot reallocation.
    - Disable time synchronization.
    - Enable lot generation at the source.
    - Enable lot movements from the queue if equipment is available for processing.
    - Enable lot movement when process time at the equipment is over.
    - Increase the simulation time internally.

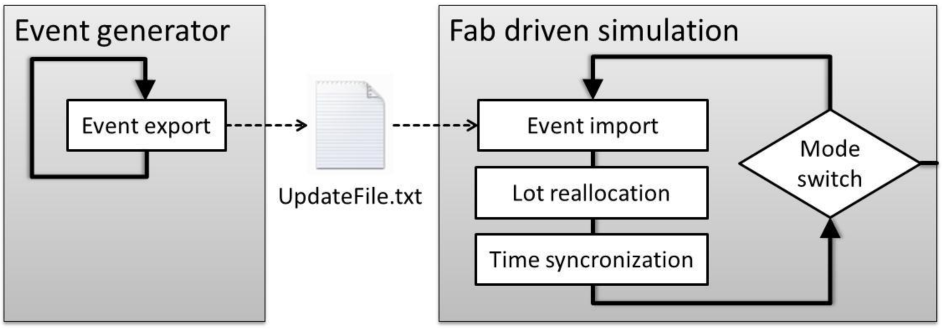


Figure 148: Simplified module execution flow chart for fab driven simulation

8.5.1 Implementation Tools

The decision is to carry out the implementation with AnyLogic 6.5.0. The relevant model extension modules are implemented in Java. The reasons for that decision are the following:

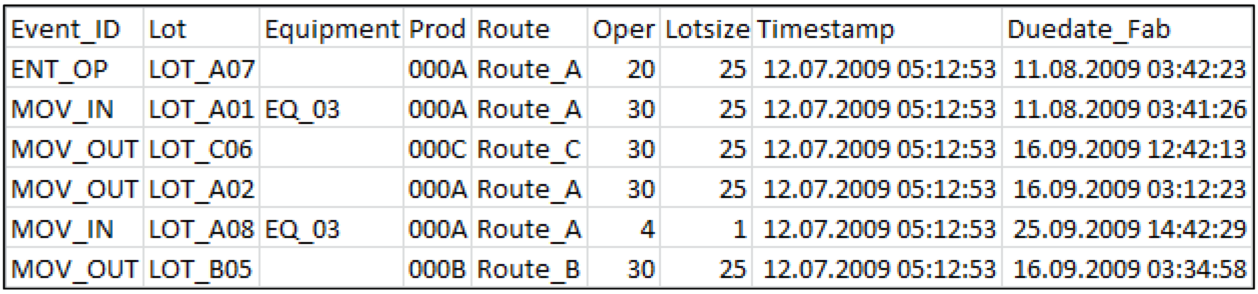
- AnyLogic has a graphical user interface, where simulation objects like queue and processor are visible with the related statistics. This is a big advantage because the concept becomes visible to the audience.
- AnyLogic is very accessible for programming purposes. Several event types like the model initialization event, the lot entry trigger, lot exit trigger are able to execute java source code. Therefore it becomes easy to extend a model with the fab driven simulation approach.

The model in AnyLogic is used as a proof of concept and as a demonstration model. It contains one work center and not the full scale fab model with all the modeling features.

For the event generation the update file is generated manually and automatically. For automated generation Real Time Dispatcher (RTD) version 7.2.4 is used to export the events. The reason is that most wafer fabs are using RTD products for dispatching. The RTD repositories hold data about the most current fab state. Event information about lot movements is available without further delay. RTD will generate text files to initialize the model, to provide the model with the current time of the wafer fab and it will provide the model with fab events like lot movements.

8.5.2 Update File Format

The update file is the link between the fab operations and the simulation run. It is specified in Table 47, including self-generated data examples. The column “Event ID” defines the type of the lot movement event. The keyword “ENT_OP” indicates that a lot enters the queue of the current operation. The keyword “MOV_IN” indicates that a lot enters a tool. The “MOV_OUT” event tells the system that the lot leaves the tool. The column “Equipment” is required only for lots with a “MOV_IN” event. In the update file, the column “Lot” contains the lot name. It must be unique because a single lot cannot have more than one update event at the same time. Additional columns like “Product”, “Route”, or “Lotsize” are used to show that it is feasible to update several other lot attributes as well.



Event_ID	Lot	Equipment	Prod	Route	Oper	Lotsize	Timestamp	Duedate_Fab
ENT_OP	LOT_A07		000A	Route_A	20	25	12.07.2009 05:12:53	11.08.2009 03:42:23
MOV_IN	LOT_A01	EQ_03	000A	Route_A	30	25	12.07.2009 05:12:53	11.08.2009 03:41:26
MOV_OUT	LOT_C06		000C	Route_C	30	25	12.07.2009 05:12:53	16.09.2009 12:42:13
MOV_OUT	LOT_A02		000A	Route_A	30	25	12.07.2009 05:12:53	16.09.2009 03:12:23
MOV_IN	LOT_A08	EQ_03	000A	Route_A	4	1	12.07.2009 05:12:53	25.09.2009 14:42:29
MOV_OUT	LOT_B05		000B	Route_B	30	25	12.07.2009 05:12:53	16.09.2009 03:34:58

Table 47: Update file example

8.6 Model

To demonstrate how the procedure works a work center model is used as an example. The current model state is defined as depicted in Figure 149. The model consists of one work center from the wafer fab. It contains one source, one sink, the equipment, and the connection between these elements. At the beginning, three lots are in the queue. Two lots are in the first equipment. One lot is situated in the second equipment.

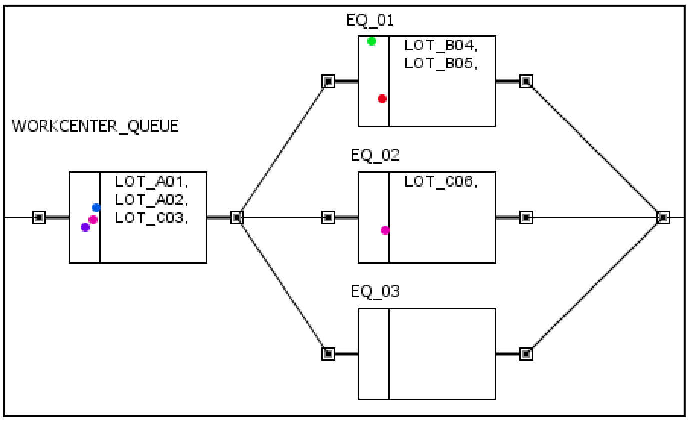


Figure 149: Model state before an update

Now a single update as depicted in Table 47 will be applied to the model in Figure 149. The lot “Lot_A07” from Table 3 enters the queue. The lot “Lot_A01” moves from the queue to equipment “EQ01”. The lot “Lot_A06” leaves the tool “EQ_01”. It will be removed by the sink. After all events have been applied, the next model state is reached, as depicted in Figure 150. These events update the old model state to the new model state. By performing such single updates in an iterative way, the lot moves in the simulation model are updated over a long time period.

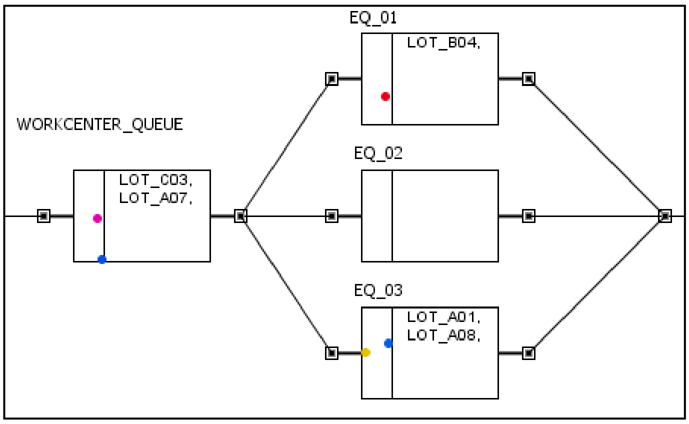


Figure 150: Model state after an update

The Real Time Dispatcher (RTD) version 7.2.4 is used to export the update file. The simulation model itself is created with AnyLogic 6.5.0. The relevant model extension modules are implemented in Java.

8.7 Results

The objective of the results section is to show the feasibility of the fab driven simulation. The first part presents the relevant test cases. The second part contains results from a fab driven simulation run in a real wafer fab.

8.7.1 Test Cases

Besides several unit tests where each module has been tested separately, numerous black box tests have been performed. The purpose of these black box tests is to check if the whole software solution is working correctly. All these test cases have been evaluated manually. The data source for the fab driven simulation is a self-generated update file.

Table 48 contains a set of the test cases that have been executed. The test cases consider the origin of the lot, the information if the lot is an existing lot from the queue or from equipment, or if this lot does not exist in the model yet. It has been checked whether the update behavior is correct, even if the previous and the new location of the lot are the same. Furthermore, several test cases have been executed for single lots, and for multiple lots that move simultaneously.

Test case description	Expected results	Test passed
A new lot enters a queue	The lot appears in the queue.	Yes
A queue lot enters a queue	The lot is in the queue.	Yes
An equipment lot enters a queue	The lot appears in the queue and disappears in the equipment.	Yes
A new lot enters an equipment	The lot appears in the equipment.	Yes
A queue lot enters an equipment	The lot appears in the equipment and disappears in the queue.	Yes
An equipment lot enters a different equipment	The lot appears in the target equipment and disappears in the previous equipment.	Yes
An equipment lot enters the same equipment	The lot is in the equipment.	Yes
A new lot moves out	This will be interpreted as data error because a lot which is not in the operation before cannot move out. It will be ignored.	Yes
A queue lot moves out	The lot appears in the sink and disappears in the queue.	Yes
An equipment lot moves out	The lot appears in the sink and disappears in the equipment.	Yes
Update the simulation model time when real time is later than simulation time	The simulation model time becomes the real time.	Yes
Update the simulation model time when real time is earlier than simulation time	The simulation model time does not change.	Yes

Table 48: Performed Test Cases

As an example following issues have been identified:

- A lot appears twice at relocation. The update method inserts the lot but it does not properly remove this lot. The bug in the removal method has been fixed properly.
- Duplicated lot entries in the event update file cause race conditions (Tanenbaum 1992). When two events exist in the update file for the same lot at the same time, the last event in the update file overwrites the first event update for the same lot. The final result depends on the sequence of both updates. The solution is to add a “single lot” constraint to the data input. So, duplicated entries are not allowed in the update at the same point in time.
- Several listening subscriptions from queue successors cause exceptions. This happened by taking a lot out of the queue without eliminating these subscriptions. This is an internal function which highly depends on the resource allocation of the simulation tool. In this case, a lot in a queue adds several subscriptions to let the follow up equipment know that when equipment capacity becomes available, the equipment can take out this particular lot. To solve this issue, these subscriptions are removed.

By performing a detailed testing incorrect behavior has been identified. All issues have been resolved afterwards. Finally the simulation model passes all test cases successfully.

8.7.2 Test at a Real Wafer Fab

Other test cases were performed in a real wafer fab. Therefore one work center has been selected to demonstrate that fab driven simulation is working under real world conditions. The criteria to select this work center are the following:

- The number of work center moves is large. Therefore it is possible to show lot movements even in a live demonstration.
- The work center is isolated. This means the lots running on this work center cannot be processed by an alternative work center.
- The work center contains only a few tools to display all of them on one screen.

An isolated work center with 5 tools has been selected. The throughput of this small work center is huge, with about 120 lots per day. This is one lot movement every 12 minutes where one move consists of “enter operation”, “move in”, and “move out”. The result of this test is depicted in Figure 151. The number of WIP lots is fluctuating over time. The runtime is about 90 minutes or 5400 seconds, where the time unit is in seconds since simulation start.

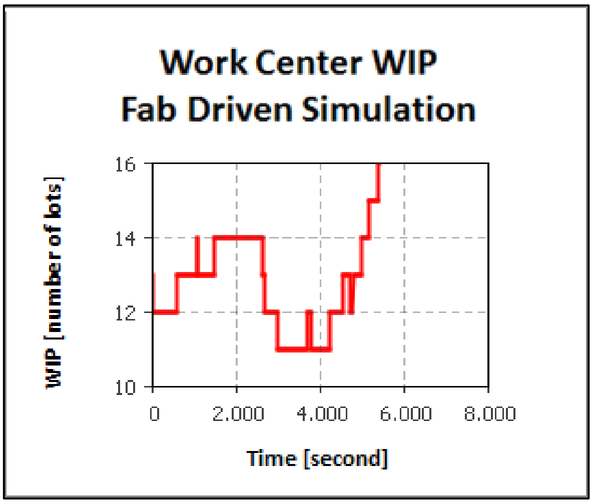


Figure 151: WIP for Fab driven simulation

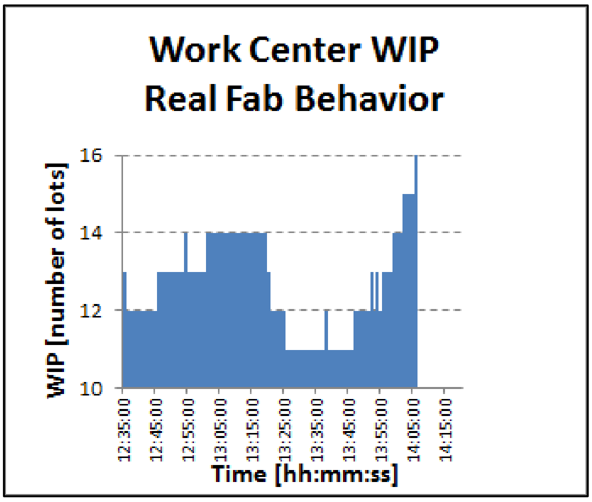


Figure 152: WIP in reality

To show that the WIP of the fab driven simulation in Figure 151 is correct, the real WIP is depicted in Figure 152. This curve shows the number of lots for the same work center at the same time period. The WIP over time is exactly the same. The simulation model update works under real world conditions within a long time period. It demonstrates that fab driven simulation is feasible.

8.8 Challenges in for Fab Driven Simulation

The challenge for fab driven simulation is to handle the large complexity of fab driven simulation. The interaction of fab events and simulation elements increases the complexity of fab driven simulation very much. The software engineering part is a highly important aspect to realize fab driven simulation for short term simulation.

From a software engineering perspective the limitation also exists, that the fab driven simulation is closely linked to the simulation tool. It is feasible to transfer the capability of fab driven simulation to another simulation tool from a different vendor easily.

The effort to create fab driven simulation for operational performance prediction is also tremendous. To illustrate this challenge, see Figure 153, 154, and 155. In semiconductor industry it is common to use fab simulation for planning purposes. Therefore it is most common to create a steady state model manually as depicted in Figure 153. The model contains data to simulate the long term behavior of the fab. It is not necessary to do a proper initialization with the current fab state. The warm-up period is not in the range of interest. Additionally a fab driven simulation approach is not useful.

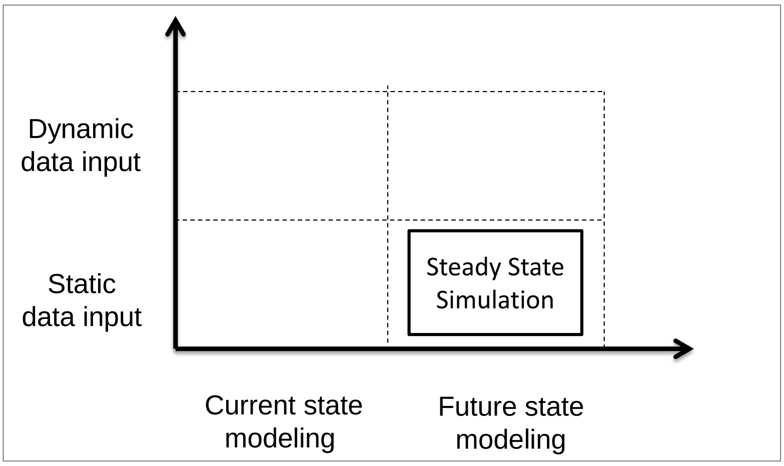


Figure 153: Simulation elements for a steady state model

When creating a short term online simulation for operational decision support, it is required to initialize the fab adequately. Therefore it is necessary to model the current fab state and the future fab state (Figure 154).

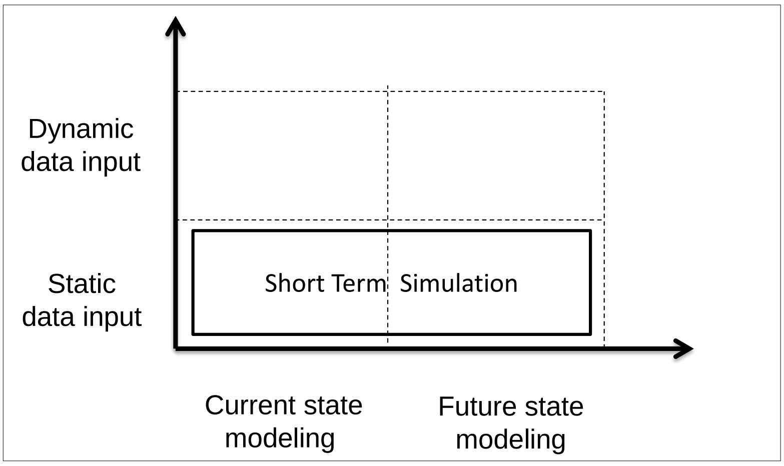


Figure 154: Simulation elements for a short term simulation model

When building the fab driven simulation approach on top of the short term simulation model, it is required to initialize the model once (static data input) and update fab events concurrently (dynamic data input). For the dynamic data input both data sets, the current state data and the future state data, also demand updates.

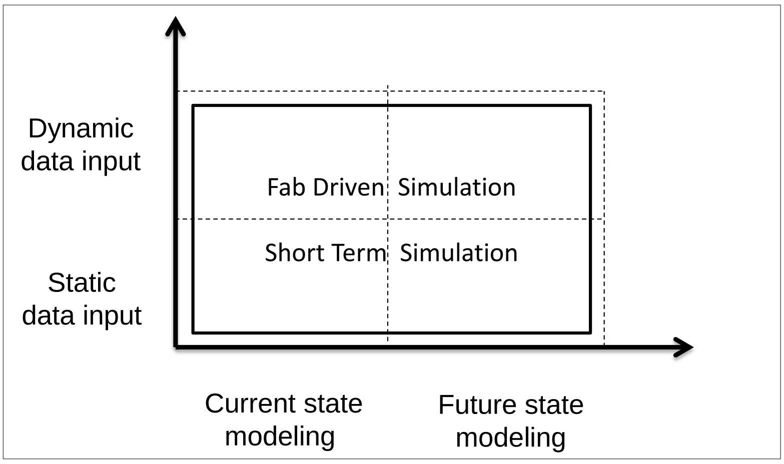


Figure 155: Simulation elements for fab driven simulation

As an example, the equipment state is used, especially for preventive maintenance (PM). For steady state simulation it is useful to model equipment downtime with the downtime distributions, using MTTF and MTTR. For short term simulation it is applicable to add the current equipment state to have an effective model initialization. For future PM modeling it is possible to utilize MTTF and MTTR as well or the PM plan, if it is available. For fab driven simulation it is required to update the current state and future states dynamically. If the current equipment state and the PM plan are changing, this needs to be adapted in the model as well. This example provides an impression of the effort to realize fab driven simulation. It can be seen that the effort for online simulation is much higher compared to steady state modeling.

9 Conclusions

The conclusions section describes the achievements of this thesis. The conclusion section is divided into the four major parts, which are the implementation of online simulation, the real time capability of fab driven simulation, the achieved forecast accuracy, and the applications of the online simulation results.

9.1 Implementation of Online Simulation in a Wafer Fab

The key advantage is that online simulation for forecasting has been successfully applied in industry. A full scale fab model generates an operational performance prediction for a short time horizon. The model generator creates the simulation model automatically. The data sources are the fab data bases. Only a few data tables are manually maintained. It takes several minutes to process the data and to generate the simulation results. The model has a high level of detail and an accurate initial state. For this thesis, it is in the range of interest to point out how exactly it is possible to implement online simulation for a full fab model. Chapter 4 and 5 describe the methodology in detail.

From simulation modeling perspective the requirements, the level of detail, simulation model features, and the validation process have been described. The major aspect is the model initialization. Numerous features enhance the model quality at simulation start.

A major aspect of online simulation is the data integration process. The full data integration of all work center information, lot information, and process information is required to reflect all fab relations in the simulation model. The data schema, the data model components, the data problems, and the solutions are described in detail. In database literature, the data integration aspect is obtainable but not dealt within the specific context of online simulation in semiconductor manufacturing. The contribution to science is a large set of simulation specific data integration conflicts and the related approach how to solve the individual conflicts.

A major failure of this thesis is the integration of the PM plans. Intensive effort has been made to collect, to process, and to integrate PM plans from different departments into a common format. The expectation is high, to increase the work center throughput accuracy by using PM plans. The analysis of the PM information shows that this information is not useful because the data quality is not sufficient. Less than 10% of the PM events are available. To model preventive maintenance (PM) in simulation, a distribution is used, similar to the unscheduled down modeling.

The recommendation for future research is to find common standards to measure the data quality for the purpose of online simulation. The first step to solve a problem is to unhide it. Therefore it is necessary to explore inconsistencies, incomplete, and incorrect data, before an online simulation project starts. The benefit of good data quality must be apparent for anyone involved in the data handling process.

From simulation modeling perspective, the automated model validation is in the range of interest for future research. It is highly useful to provide feedback to the user if the model generation fails, if the model results differ from reality, and if the model differs from previous forecast models. A highly automated monitoring and improvement is essential to keep the forecast quality on a high level.

9.2 Fab Driven Simulation

One limitation to achieve real time capability is the time consuming data extraction, data cleaning, and the model initialization process. The fab driven simulation has been developed to overcome those limitations.

This thesis shows that fab driven simulation is feasible. A prototype has been implemented to update the lot moves of one work center. This prototype demonstrates that it is possible to continuously update a simulation model during runtime. By having an updated model, the user is able to run the simulation immediately. The big advantage of fab driven simulation is that it is not necessary to wait for data processing and model initialization before the simulation model is executed. Compared to the described approaches in literature, the fab driven simulation contributes significantly to achieve real time capability of an online simulation model.

The experience from this project is that the complexity of fab driven simulation is on a high level, especially the implementation. The interaction of fab events and simulation elements increases the complexity of fab driven simulation significantly. The software engineering part is a highly important aspect to realize fab driven simulation for short-term simulation. Furthermore there is a limitation from the software engineering perspective. The fab driven simulation feature is closely tied to the simulation tool. So it is not easy to transfer the capability of fab driven simulation to another simulation tool from a different vendor.

A key challenge for future research is to keep extending the fab driven simulation approach. First, it is required to increase the scale of the model to the whole wafer fab. Secondly, it is necessary to update more simulation event types besides the lot moves. The third element is to clone the fab driven simulation process during runtime. The objective is to let the fab driven simulation model run continuously. A trigger copies this process. This copied process is used to create a short-term forecast, while the original process keeps on synchronizing with the real fab.

9.3 Accuracy

For this thesis the focus is to show the achieved accuracy of online simulation. The analysis includes the influence of the stochastic effects and the influences of the model initialization.

The accuracy has been analyzed for different level of detail, from fab level via work center level down to lot level. For each level of detail, the relevant KPI’s have been selected to measure the forecast accuracy. On fab level the forecast error is very small. The WIP deviation (MAPE) is around 2%. On lot level the deviation is very high as expected. For a 7 day forecast, the cycle time deviation already reaches a level of 1.7 days. Therefore a lot based arrival forecast is not feasible. On work center level the WIP forecast is not useable but the lot arrival forecast shows good results. The forecast error for work center arrivals is only 11% for those work centers with at least 1% fab moves. For a mature logic fab with high variability these results are good. The forecast results are usable to increase the fab performance.

A single diagram shows the relations between forecast error, forecast horizon and the level of detail. The relationships between forecast error, forecast horizon and the level of detail have been analyzed. Regarding the forecast horizon, the model behavior on fab level is stable. This behavior is similar to reality. The accuracy on fab level is very high. The work center arrivals also show good results for a forecast period of more than 7 days. The lot level has the highest level of detail. It is considerably affected by uncertainties. The deviation from reality is massive, even at simulation start. Even for a forecast horizon of less than one day, the lot arrival forecast deviates much.

The sensitivity analysis shows the effect of the model initialization elements. The initialization of WIP lots with the current operation is essential. Except for the initialization with the current operation, the long term effect of a detailed initialization is very small. A significant effect is visible for the first day only. The major increase of the forecast accuracy for the first day is caused by the assignment of the current station and the remaining process time. The recommendation is to initialize the fab with a high level of detail, only if the desired forecast horizon is one day or less.

For future work it is important to continue enhancing the model accuracy. Efforts have been made in the past to handle the data issues and the uncertainties. Nevertheless unresolved issues still exist. Additional work is necessary for model improvement and model validation. It is required to increase the degree of automation for the validation process itself. To increase the model accuracy, another aspect is to further increase the data quality directly in the fab data sources.

For future research of online simulation forecasting, it is recommended to publish the forecast error for several KPIs. It is useful to compare the results and to gather information about for which KPI and which level of detail a forecast is applicable. The target is a comparison of forecast methods in combination with specific fab environments, like high mix/low volume wafer fabs. The comparison of different forecasting methods reveals advantages and disadvantages of each method.

9.4 Applications

The question for this thesis is how exactly online simulation enhances the fab performance. It is necessary to evaluate if the applications are compatible with simulation forecast to reach that goal. Therefore one question is to identify if online simulation satisfies the requirements of an application.

To answer these questions, several operational problems (or applications) have been identified. Short term forecasting is capable to support these applications with future fab information. This thesis contains a list with several details of potential applications. The process description is available to identify the relevant applications. For online simulation the feasible applications are PM planning, backup tool activation, dedication, and dispatch rule selection. They are capable of improving the fab performance. The work center lot arrival forecast, with a forecast error of less than 11%, provides the input data for these applications. A detailed example also shows the process, how online simulation is capable of improving fab performance. A potential solution to reduce the peak cycle time of 38% and the peak WIP of 25% has been identified for one work center.

Glossary

Term	Description
ASAP	AutoSched AP, Simulation Software
CT	Cycle time
DWH	Data warehousing
ETL	Extract, transform, and load
Forecast horizon	For simulation forecasting the forecast horizon is the simulated time period.
KPI	Key performance indicator
Lot	Product Unit. In semiconductor manufacturing a lot carries multiple Wafers.
Lot Arrivals	The event when a lot arrives at an operation.
Lot Moves	The event when a lot leaves the operation and the equipment.
ME	Mean error
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MTTF	Mean time between failure
MTTR	Mean time to repair
PM	Preventive maintenance. Scheduled equipment downtime which it is known beforehand.
RPT	Raw process time. The time where the lot is in the equipment and will be processed.
RTD	Real Time Dispatcher, Dispatching Software
Simulation execution time	The simulation execution time (or runtime) is the computation time to run execute the simulation model run and generate the results.
WIP	Work in progress

List of Publications

For this thesis the following conference papers have been published. The focus of these publications is on simulation based optimization, online simulation, data integration, and forecasting.

D. Noack, and O. Rose. 2008. A Simulation Based Optimization Algorithm for Slack Reduction and Workforce Scheduling. In Proceedings of the 2008 Winter Simulation Conference. pp. 1989-1994.

D. Noack, B. P. Gan, P. Lendermann, and O. Rose. 2008. An Optimization Framework for Waferfab Performance Enhancement. In Proceedings of the 2008 Winter Simulation Conference. pp. 2194-2200.

R. Kohn, D. Noack, M. Mosinski, Z. Zhou, and O. Rose. 2009. Evaluation of Modeling, Simulation and Optimization Approaches for Work Flow Management in Semiconductor Manufacturing. In Proceedings of the 2009 Winter Simulation Conference. pp. 1592-1600.

O. Rose, M.F. Majohr, E. Angelidis, F. S. Pappert, and D. Noack. Personaleinsatz- und Ablaufplanung für komplexe Montagelinien mit MARTA 2. In Simulation und Optimierung in Produktion und Logistik: Praxisorientierter Leitfaden mit Fallbeispielen. edited by L. März, W. Krug, O. Rose, G. Weigert. Springer. pp. 93-104.

D. Noack, R. Kohn, M. Mosinski, Z. Zhou, O. Rose, W. Scholl, P. Lendermann, and B. P. Gan. 2010. Data Modeling for Online Simulation - Requirements and Architecture. In Proceedings of the 2010 FAIM Conference. pp. 1037-1044.

W. Scholl, B. P. Gan, M. L. Peh, P. Lendermann, D. Noack, O. Rose, and P. Preuss. 2010. Towards Realization of a High-Fidelity Simulation Model for Short-Term Horizon Forecasting in Wafer Fabrication Facilities. In Proceedings of the 2010 Winter Simulation Conference. pp. 2563-2574.

D. Noack, M. Mosinski, O. Rose, P. Lendermann, and B.P. Gan. 2011. Challenges and Solution Approaches for the Online Simulation of Semiconductor Wafer Fabs. In Proceedings of the 2011 Winter Simulation Conference. pp. 1845-1856.

M. Mosinski, D. Noack, F. Pappert, O. Rose, and W. Scholl. 2011. Cluster Based Analytical Method for the Lot Delivery Forecast in Semiconductor Fab with wide Product Range. In Proceedings of the 2011 Winter Simulation Conference. pp. 1834-1844.

W. Scholl, B. P. Gan, P. Lendermann, D. Noack, O. Rose, P. Preuss, and F. S. Pappert. 2011. Implementation of a Simulation-Based Short-Term Lot Arrival Forecast in a Mature 200mm Semiconductor Fab. In Proceedings of the 2011 Winter Simulation Conference. pp. 1932- 1943.

References

W.M.P. van der Aalst. 1998. The Application of Petri Nets to Workflow Management. The Journal of Circuits, Systems and Computers. Vol.8, No 1, pp. 21-66.

J. Aitchison, and I. R. Dunsmore. 1975. Statistical Prediction Analysis. Cambridge University Press.

L. F. Atherton, and R. W. Atherton. 1995. Wafer Fabrication: Factory Performance and Analysis. Kluwer Academic Publishers.

H. Aydt. 2011. An Agent-based Symbiotic Simulation Framework for Automated Problem Solving". PhD Thesis Nanyang Technological University.

H. Aydt, S. J. Turner, W. Cai, and M. Y. H. Low. 2009. Research Issues in Symbiotic Simulation. In Proceedings of the 2009 Winter Simulation Conference. pp. 1213-1222.

H. Aydt, S. J. Turner, W. Cai, M. Y. H. Low, P. Lendermann, and B. P. Gan. 2008. Symbiotic Simulation Control in Semiconductor Manufacturing. In Proceedings of the International Conference on Computational Science. pp. 26–35.

S. Bagchi, C. H. Chen-Ritzo, S. T. Shikalgar, and M. Toner. 2008. A Full-Factory Simulator as a Daily Decision-Support Tool for 300mm Wafer Fabrication Productivity. In Proceedings of the 2008 Winter Simulation Conference. pp. 2021-2029.

O. Balci. 1986. Credibility Assessment of Simulation Results. In Proceedings of the Conference on Simulation Methodology and Validation.

O. Balci. 1998. Verification, Validation, and Accreditation. In Proceedings of the 1998 Winter Simulation Conference. pp. 135-141.

E. Böhl. 2010. Analyse und Vorhersage von Wartungen auf Grundlage historischer Daten und vorgegebener Wartungspläne. Master Thesis Dresden University of Technology.

C. S. Chong, and A. I. Sivakumar. 2003. Simulation-Based Scheduling for Dynamic Discrete Manufacturing. In Proceedings of the 2003 Winter Simulation Conference. pp. 1465-1473.

R. Crosbie. 2010. Grand Challenges in Modeling and Simulation. In Proceedings of the 2010 Grand Challenges in Modeling and Simulation Conference. The Society for Modeling and Simulation International.

W. Dangelmaier, K. R Mahajan, T. Seeger, B. Klöpper, and M. Aufenanger. 2006. Simulation Assisted Optimization and Real-Time Control Aspects of Flexible Production Systems Subject to Disturbances. In Proceedings of the 2006 Winter Simulation Conference. pp. 1785-1795.

E. V. Denardo. 2003. Dynamic Programming: Models and Application. Dover Publications.

A. Dietmair, and A. Verl. 2010. Energy Efficiency Optimization in Production Planning and Control. In Proceedings of the 20th International Conference on Flexible Automation and Intelligent Manufacturing. pp. 278- 285.

J. Domaschke, S. Brown, J. K. Robinson, and F. Leibl. 1998. Effective Implementation of Cycle Time Reduction Strategies for Semiconductor Back-End Manufacturing. In Proceedings of the 1998 Winter Simulation Conference. pp 985-992.

M. Dümmler. 1999. Using simulation and genetic algorithms to improve cluster tool performance. In Proceedings of the 1999 Winter Simulation Conference. pp 875-879.

M. Dümmler. 2004. Modeling and Optimization of Cluster Tools in Semiconductor Manufacturing. Dissertation, University of Würzburg.

G. R. Drake, and J. S. Smith. 1996. Simulation System for Real-Time Planning, Scheduling, and Control. In Proceedings of the 1996 Winter Simulation Conference. pp. 1083-1090.

J. W. Fowler, O. Rose, S. Strassburger, and S. Turner. 2002. Grand Challenges in Modeling & Simulation. Dagstuhl Seminar Nr 02351. Slides of the Manufacturing Working Group.

A. Frantsuzov. 2011. Automated Analytical Equipment Modeling in Semiconductor Manufacturing. Master Thesis Dresden University of Technology.

R. Fujimoto, D. Lunceford, E. Page, and A. M. Uhrmacher (editors). 2002. Grand Challenges for Modeling and Simulation. Dagstuhl Seminar Nr 02351. Report Nr 350.

C. Girault, and R. Valk. 2002. Petri Nets for Systems Engineering. 1st ed. Springer.

A. Hanisch,J. Tolujew, and T. Schulze. 2005. Initialization of Online Simulation Models. In Proceedings of the 2005 Winter Simulation Conference. pp. 1795-1803.

Y. He, M.C. Fu, and S.I. Marcus. 2000. Simulation-based Approach for Semiconductor Fab-Level Decision Making - Implementation Issues. Technical Report TR2000-48, Institute for Systems Research, University of Maryland.

P. E. Heegaard, and K. S. Trivedi. 2009. Survivability Modeling with Stochastic Reward Nets. In Proceedings of the 2009 Winter Simulation Conference. pp. 801-818.

K. Hoad, S. Robinson, and R. Davies. 2008. Automating Warm-Up Length Estimation. In Proceedings of the 2008 Winter Simulation Conference.

J. R. Holton. 2004. An Introduction to Dynamic Meteorology. 4th ed. Elsevie.

W. Hopp, and M. Spearman. 2000. Factory Physics, 2nd ed. McGraw-Hill/Irwin.

S. Horn. 2008. Simulationsgestützte Optimierung von Fertigungsabläufen in der Produktion elektronscher Halbleiterspeicher. PhD Thesis Dresden University of Technology.

S. Horn, G. Weigert, and S. Werner. 2005. Data Coupling Strategies in Production Environments. In 28th International Spring Seminar on Electronics Technology ISSE. pp. 278-282.

S. Horn, G. Weigert, S. Werner, and T. Jähnig. 2006. Simulation Based Scheduling System in a Semiconductor Backend Facility. In Proceedings of the 2006 Winter Simulation Conference. pp. 1741-1748.

R. J. Hyndman, and A. B Koehler. 2005. Another Look at Measures of Forecast Accuracy. International Journal of Forecasting, Volume 22, Issue 4, pp. 679-688.

S. Jensen. 2007. Eine Methodik zur Teilautomatisierten Generation von Simualtionsmodellen aus Produktionsdatensystemen am Beispiel einer Jop Shop Fertigung. PhD Thesis University of Kassel.

M. Klein, and A. Kalir. 2006. Improved Simple Simulation Models for Semiconductor Wafer Factories. In Proceedings of the 2006 Winter Simulation Conference. pp. 1708-1712.

A. Klemmt, S. Horn, and G. Weigert. 2008. Analysis and Coupling of Simulation-based Optimization and MIP Solver Methods for Scheduling of Manufacturing Processes. In Proceedings of the 2008 FAIM Conference. pp. 1228-1235.

A. Law, and W. D. Kelton. 1999. Simulation Modeling and Analysis. 3rd ed. McGraw-Hill.

W. Lehner. 2003. Datenbanktechnologie für Data-Warehouse-Systeme. dpunkt Verlag.

C. Lindemann. 1998. Performance Modelling with Deterministic and Stochastic Petri Nets. Wiley.

M. Y. H. Low, K. W. Lye, P. Lendermann, S. J. Turner, R. T. W. Chim, and S. H. Leo. 2005. An Agent-Based Approach for Managing Symbiotic Simulation of Semiconductor Assembly and Test Operation. In Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems. pp. 85–92.

L. März, W. Krug, O. Rose, and G. Weigert. 2011. Simulation und Optimierung in Produktion und Logistik. Springer.

P. S. Mahajan, and R. G. Ingalls. 2004. Evaluation of Methods Used to Detect Warm-Up Period in Steady State Simulation. In Proceedings of the 2004 Winter Simulation Conference. pp. 663-671.

S. Makridakis, S. Wheelwright, and R. J. Hyndman. 1998. Forecasting: Methods and Applications. 3rd ed. JohnWiley & Sons: New York.

S. C. Mathewson. 1984. The Application of Program Generator Software and its Extensions to Discrete-Event Simulation Modeling. IIE Transactions. Vol. 16, No. 1, pp. 3-18.

J. McGregor. 2007. The Common Platform Technology: A New Model for Semiconductor Manufacturing. Report. In-Stat.

P. McNally, and C. Heavey. 2004. Developing Simulation as a Desktop Resource. International Journal of Computer Integrated Manufacturing. Vol. 17, pp. 435-450.

P. Mertens, and S. Rässler. 2004. Prognoserechnung. 6th ed. Physica-Verlag.

L. Mönch, O. Rose, and R. Sturm. 2003. A Simulation Framework for the Performance Assessment of Shop-Floor Control System. Simulation : Transactions of the Society for Modeling and Simulation International. Vol. 79, Issue 3, pp. 163-170.

J. R. Morrison. 2011. On the Fidelity of the Ax+B Equipment Model for Clustered Photolithography Scanners in Fab-Level Simulation. In Proceedings of the 2011 Winter Simulation Conference. pp. 2034-2044.

H. Müller Sommer. 2012. Wirtschaftliche Generierung von Belieferungssimulationen unter Verwendung rechnerunterstützter Plausibilisierungsmethoden für die Bewertung der Eingangsdaten. PhD Thesis Ilmenau University of Technology.

K. Neuman, and M. Morlock. 2002. Operations Research, Carl Hanser Verlag München Wien.

H. Niedermayer, and O. Rose. 2003. A Simulation-Based Analysis of the Cycle Time of Cluster Tools in Semiconductor Manufacturing. In Proceedings of the 15th European Simulation Symposium. pp. 26-29.

J. R. Norris. 1998. Markov Chains. Cambridge University Press.

P.M. Oldfather, A.S. Ginsberg, and H.M. Markowitz. 1966. Programming by Questionnaire: How to Construct a Program Generator. Rand Report.

J. Pearl. 1984. Heuristics: Intelligent Search Strategies for Computer Problem Solving. Addison Wesley Longman Publishing Co.

T. Phillips. 1998. Autosched AP by Autosimulations. In Proceedings of the 1998 Winter Simulation Conference. pp. 219-222.

M. Pinedo. 2002. Scheduling Theory, Algorithms, and Systems. 2nd ed. Prentice Hall.

L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data Quality Assessment. Communications of the ACM. Vol. 45, No. 4. pp. 211-218.

J. Potoradi, O.S. Boon, S. J. Mason, J. W. Fowler, and M. E. Pfund. 2002. Using Simulation-Based Scheduling to Maximize Demand Fulfillment in a Semiconductor Assembly Facility. In Proceedings of the 2002 Winter Simulation Conference. pp. 1857-1861.

M. Rabe, S. Spieckermann, and S. Wenzel. 2007. Verifikation und Validierung für die Simulation in Produktion und Logistik. Springer.

S. Radloff, M. Abravanel, B. Rhoads, D. Steeg, P. van der Meulen, and M. Petraitis. 2009. First Wafer Delay and Setup: How to Measure, Define and Improve First Wafer Delays and Setup Times in Semiconductor Fabs. In Proceedings of the 2009 IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC). pp. 86-90.

E. Rahm, and P. A. Bernstein. 2001. A Survey of Approaches to Automatic Schema Matching. The VLDB Journal. Vol. 10, No. 4, pp. 334-350.

E. Rahm, and H. H. Do. 2000. Data Cleaning: Problems and Current Approaches. Bulletin of the Technical Committee on Data Engineering, Vol. 23, No. 4, pp 3-13.

L. G. Randell, G. S. Bolmsjö. 2001. Database Driven Factory Simulation: A Proof-Of-Concept Demonstrator. In Proceedings of the 2001 Winter Simulation Conference. pp 977-983.

H.A. Reijers, and W.M.P. van der Aalst. 1999. Short-Term Simulation: Bridging the Gap Between Operational Control and Strategic Decision Making. In Proceedings of the IASTED International Conference on Modeling and Simulation. pp. 417-421.

W. Reisig. 1985. Petri nets: An Introduction. Springer.

N. Robertson, and T. Perera. 2002. Automated Data Collection for Simulation. Simulation Practice and Theory. Vol. 9, pp. 349-364.

J. K. Robinson. 1998. Capacity Planning in a Semiconductor Wafer Fabrication Facility with Time Constraints between Process Steps. PhD Thesis University of Massachusetts.

S. Robinson. 2002. A Statistical Process Control Approach for Estimating the Warm-Up Period. In Proceedings of the 2002 Winter Simulation Conference. pp. 532-540.

S. Robinson. 2004. Simulation: The practice of model development and use. John Wiley & Sons.

O. Rose. 2006 (a). Analysis of a Semiconductor Wafer Fab with an Intermediate Storage for Partially Finished Logic Products. In Proceedings of the Industrial Simulation Conference. pp. 319-324.

O. Rose. 2006 (b). Implementation of a Simulation-Based Optimizer for Semiconductor Wafer Factories. IEEE Conference on Emerging Technologies and Factory Automation, ETFA '06. pp. 943-949.

C. Roser, M. Nakano, and M. Tanaka. 2001. A Practical Bottleneck Detection Method. In Proceedings of the 2001 Winter Simulation Conference. pp. 949-953.

R.G. Sargent. 2005. Verification and Validation of Simulation Models. In Proceedings of the 2005 Winter Simulation Conference. pp. 130-143.

W. Scholl. 2008. Coping with Typical Unpredictable Incidents in a Logic Fab. In Proceedings of the 2008 Winter Simulation Conference. pp. 2030-2034.

L. Schruben, H. Singh, and L. Tierney. 1983. Optimial Tests for Initialization Bias in Simulation Output. Operations Research Vol. 31 No. 6.

C. J. Schuster. 2003. No-wait Job-Shop-Scheduling: Komplexität und Local Search. PhD Thesis University of Duisburg-Essen.

SEMI. 2003 (a). Provisional Guide for Definition and Calculation of Overall Factory Efficiency (OFE) and other Associated Factory-Level Productivity Metrics. SEMI E124-1103.

SEMI. 2003 (b). Specification for Definition and Measurement of Equipment Reliability, Availability, and Maintainability (RAM). SEMI E10-0304.

A. I. Sivakumar, and C. S. Chong. 2001. A Simulation Based Analysis of Cycle Time Distribution, and Throughput in Semiconductor Backend Manufacturing. Computers in Industry. Vol. 45. Issue 1. pp. 59–78. Elsevier.

A. I. Sivakumar, C. Qi, and A. K. H. Darwin. 2008. Experimental Study on Variations of WIPLOAD Control in Semiconductor Wafer Fabrication Environment. In Proceedings of the 2008 Winter Simulation Conference. pp. 2035-2040.

A. Skoogh, and B. Johansson. 2008. A Methodology for Input Data Management in Discrete Event Simulation Projects. In Proceedings of the 2008 Winter Simulation Conference. pp. 1727-1735.

A. Skoogh, J. Michaloski and N. Bengtsson. 2011. Towards Continuously Updated Simulation Models: Combining Automated Raw Data Collection and Automated Data Processing. In Proceedings of the 2008 Winter Simulation Conference. pp 1678-1689.

A. Skoogh. 2011. Automation of Input Data Management. Dissertation, Chalmers University of Technology, Gothenburg.

J. S. Smith, R. A. Wysk , D. T. Sturrock, S. E. Ramaswamy, G. D. Smith, and S. B. Joshi. 1994. Discrete Event Simulation for Shop Floor Control. In Proceedings of the 1994 Winter Simulation Conference. pp. 962-969.

Y. J. Son, and R. A. Wysk. 2001. Automatic Simulation Model Generation for Simulation-Based, Real-Time Shop Floor Control. In Computers in Industry. Vol. 45, pp. 291–308. Elsevier.

A.S. Tanenbaum. 1992. Modern Operating Systems. Prentice Hall.

D. S. Walonick. 1993. An Overview of Forecasting Methodology. http://www.statpac.org/research-library/forecasting.htm [accessed 24.1.2012].

T. W. Wang, and K. E. Murphy. 2004. Semantic Heterogeneity in Multidatabase Systems: A Review and a Proposed Meta-Data Structure. Journal of Database Management, Vol. 15, No. 4, pp. 71-87.

J. H. Wilson, and B. Keating. 1994. Business Forecasting. 2nd ed. Irvin.

J. R. Wilson, and A. A. B. Pritsker. 1978. A survey of research on the simulation startup problem. Simulation Vol. 31 No. 2.

S. D. Wu, and R. A. Wysk. 1989. An Application of Discrete-Event Simulation to Online Control and Scheduling of Flexible Manufacturing. International Journal of Production Research. Vol. 27, No. 9.

C.Y. Yu, and H.P. Huang. 2002. On-Line Learning Delivery Decision Support System for Highly Product Mixed Semiconductor Foundry. IEEE Transactions on Semiconductor Manufacturing. Vol. 15, No. 2, pp. 274-278.

H. Zhang, Z. Jiang, and C. Guo. 2008. Simulation-Based Optimization of Dispatching Rules for Semiconductor Wafer Fabrication System by the Response Surface Methodology Scheduling. The International Journal of Advanced Manufacturing Technology. Vol. 41, Nr. 1 pp 101-121. Springer.

H. Zisgen, I. Meents, B. R. Wheeler, and T. Hanschke. 2008. A Queueing Network Based System to Model Capacity and Cycle Time for Semiconductor Fabrication. In Proceedings of the 2008 Winter Simulation Conference. pp. 2067-2074.

H. Zisgen, and B. R. Wheeler. 2007. WIP Movement Prediction by EPOS in IBM's 300mm fab. 4th ISMI Symposium on Manufacturing Effectiveness. Sematech.