ARCHITECTURE FOR SIMULATION-BASED PERFORMANCE ASSESSMENT OF PLANNING APPROACHES IN SEMICONDUCTOR MANUFACTURING

Thomas Ponsignon\textsuperscript{1,2} \hspace{1cm} Lars Mönch\textsuperscript{2}

\textsuperscript{1}Infineon Technologies AG  \hspace{1cm} \textsuperscript{2}Dept. of Mathematics and Computer Science
Am Campeon 1-12  \hspace{1cm} University of Hagen
Neubiberg, 85579, GERMANY  \hspace{1cm} Hagen, 58097, GERMANY

ABSTRACT

Complex manufacturing systems, such as wafer fabrication facilities (wafer fabs), are characterized by a diverse product mix that is changing over time, re-entrant process flows due to expensive machinery, different process types, and different kinds of internal and external disruptions. In this context, simulation-based architectures have been extensively used to assess the performance of different production control approaches (cf. Mönch 2007). For example, scheduling approaches in a rolling horizon setting are considered. However, the interaction between the production planning and the control process is predominantly neglected so far. Hence, in most cases the performance of planning algorithms, also in Enterprise Resource Planning (ERP) and Advanced Planning and Scheduling (APS) systems, is assessed based on the analysis of single test instances, or a performance assessment is even not carried out. Thus, in this paper it is shown how a simulation-based architecture that allows modeling the typical stochastic and dynamic behavior of a complex manufacturing system can be used to assess the performance of production planning approaches.

The paper is organized as follows. In Section 2, we describe the problem under consideration, and we discuss related literature. The respective elements of the proposed architecture are presented as well as its implementation and application in Section 3. A performance assessment methodology is also outlined. Then, we assess the performance of two master planning schemes for a wafer fab in Section 4. We also explain how the simulation model is reduced to decrease the computational burden. Finally, we present some conclusions and future research directions in Section 5.

1 INTRODUCTION

Complex manufacturing systems, such as wafer fabs, are characterized by a diverse product mix that is changing over time, re-entrant process flows due to expensive machinery, different process types, and different kinds of internal and external disruptions. In this context, simulation-based architectures have been extensively used to assess the performance of different production control approaches (cf. Mönch 2007). For example, scheduling approaches in a rolling horizon setting are considered. However, the interaction between the production planning and the control process is predominantly neglected so far. Hence, in most cases the performance of planning algorithms, also in Enterprise Resource Planning (ERP) and Advanced Planning and Scheduling (APS) systems, is assessed based on the analysis of single test instances, or a performance assessment is even not carried out. Thus, in this paper it is shown how a simulation-based architecture that allows modeling the typical stochastic and dynamic behavior of a complex manufacturing system can be used to assess the performance of production planning approaches.

The paper is organized as follows. In Section 2, we describe the problem under consideration, and we discuss related literature. The respective elements of the proposed architecture are presented as well as its implementation and application in Section 3. A performance assessment methodology is also outlined. Then, we assess the performance of two master planning schemes for a wafer fab in Section 4. We also explain how the simulation model is reduced to decrease the computational burden. Finally, we present some conclusions and future research directions in Section 5.

2 PROBLEM DESCRIPTION AND RELATED LITERATURE

In this section, we introduce some insights from systems theory to describe the researched problem. Then, we discuss related literature.
2.1 Insights from Systems Theory and Problem Description

In systems theory, we differentiate between base system and base process. The base system \( B \) comprises objects from the manufacturing system that represent the existing resources, i.e., machines, whereas the base process \( BP \) states how the resources are used by specifying process flows. In addition, a control process \( CP \) is used to determine instructions \( mc \) for the base process. Note that control decisions can only affect objects that are already in the base system. The control process is executed by the control system \( C \). Moreover, a planning process \( PP \) establishes which orders have to be started in which period of time. The corresponding instructions are denoted by \( mp \). They are provided to the control process to launch orders into the base system. The planning is performed by the planning system \( P \). It also considers objects that are not yet released into the base system. The interaction between the planning system, the control system, and the base system is summarized on Figure 1.

![Figure 1: Interactions between planning, control, and base systems](image)

In this paper, we aim to assess the performance of specified planning algorithms \( PA_i \) that are used within planning processes \( PP_i \). The performance measure values are determined by the base system and the base process. As a prerequisite, we assume that the control algorithm \( CA_0 \) and the control process \( CP_0 \) are given.

2.2 Related Literature

In this section, we discuss related literature. We distinguish between work that assess the planning performance by considering single test instances and other studies that use simulation to establish a rolling horizon environment. Bermon and Hood (1999), Barahona et al. (2005), and Zobolas, Tarantilis, and Ioannou (2008) belong to the first category. The limitation of this method is that dynamic and stochastic characteristics of the base system and process are not appropriately captured.

Planning processes are often analyzed in the context of supply chain management by means of simulation models. For this purpose, System Dynamics (Kleijnen 2005) and discrete-event simulation (Horiguchi et al. 2001; Chong et al. 2006) are widely used as simulation techniques. Indeed both methods allow modeling the time-dependent behavior of the supply chain. In addition, Venkateswaran, Son, and Jones (2004) introduce a hierarchical production planning approach that explicitly differentiates between planning and control levels. These levels are modeled using System Dynamics and discrete-event simulation, respectively, while the High-Level-Architecture (HLA) ensures the synchronization between both simulation models. Nevertheless, none of these papers show a dedicated architecture that interlinks the planning system, the control system, and the base system since the algorithms are directly implemented in the simulation software. Mönch, Rose, and Sturm (2003) and Mönch (2007) suggest a simulation-based architecture for performance assessment of production control approaches. The core of the developed archi-
Ponsignon and Mönch

Architecture is a blackboard-like data layer used as an interface between the control system and the base system. The base system is represented by an appropriate simulation model.

Related literature lacks an appropriate architecture devoted to performance assessment that simultaneously considers production planning and control approaches. In this paper, we further customize and extend the architecture described by the second present author in (Mönch 2008) to bridge this gap.

3 ARCHITECTURE FOR SIMULATION-BASED PERFORMANCE ASSESSMENT OF PLANNING APPROACHES

In this section, the respective elements of the proposed architecture are described. Then, its implementation and application are outlined. Finally, it is shown how the developed architecture is used to assess the performance of planning algorithms based on the approach in Mönch (2007) for production control schemes.

3.1 Description of the Architecture

As depicted in Figure 2, the proposed architecture consists of a production planning system whose performance has to be analyzed, a given production control system, a simulation model that represents the execution level, i.e., base system and process, a blackboard-like data layer, a demand generator that provides customer orders and additional forecast, and a performance assessment module.

![Figure 2: Architecture for simulation-based performance assessment of planning approaches](image)

The core of the proposed architecture is the blackboard-like data layer that is used as an interface between the simulation model and the planning and control modules. It comprises a mirror image of the objects of the base system and process (Mönch, Rose, and Sturm 2003). The status of these objects is updated with the help of event-driven notifications from the simulator. For instance, as soon as a lot completes a processing step and moves to the next one, its position on the process flow is updated in the data layer. Thus, it allows providing the current state of the system to the production planning and control process. The data layer mimics the data that is found in operational information systems of the wafer fab. Therefore, it includes processing steps, process flows, offered capacities, bottleneck information, product aggregation structure, and orders.

A demand generator is required in the architecture to feed the production planning algorithms. Since market requirements are not entirely known while planning is done a couple of weeks or months ahead, we have to distinguish between customer orders and additional forecast that are an input of the order management process and the sales and operations planning, respectively. The generated demand is stored
in the data layer. This allows for the use of statistical methods to determine additional forecast. Finally, the architecture also includes a module dedicated to performance assessment.

### 3.2 Implementation of the Architecture

The base system is implemented using the AutoSched AP simulator, a class library that offers customization functionalities using the C++ programming language. The data model implemented in the blackboard-like data layer is an extended and refined version of the model proposed in Mönch (2008). It comprises objects both from the base system and process as well as some additional classes with aggregated data. The blackboard-like data layer is coded in C++.

### 3.3 Application of the Architecture

Algorithm 1 shows the scheme used for simulation-based performance assessment of planning algorithms in pseudo-code notation. The repeat loop allows simulating the base process over the entire horizon. However, the simulation engine is regularly stopped after a given amount of time to perform the next run of the planning algorithm. As soon as the control is given to the planning system, statistics from the previous simulation period, i.e., throughput, work-in-process, cycle times, and waiting times, are stored in the performance assessment module. Then, current customer orders and additional forecast are provided by the demand generator to the data layer and the planning algorithm.

Moreover, the actual state of the base system and process, i.e., work-in-process, is communicated to the planning algorithm. Then, the planning algorithm is performed. As a result, we obtain production requests with due dates that are transferred to the production control system. The production control functionality offered by the proposed architecture is rather simple. It basically splits the production requests into lots to be started based on the standard lot size, and it assigns start dates to the lots using a simple backwards termination scheme based on the average product cycle times. The lots to be released are held in an order pool, and the control is given back to the simulator. Whenever a release date of a lot is reached during the simulation, a new lot object is created in the data layer, and its status is updated accordingly to its progress. We use simple dispatching rules like First-In-First-Out (FIFO) to dispatch the lots within the wafer fab. Finally, the performance assessment module determines performance measure values.

### 3.4 Performance Assessment Methodology

The production planning algorithm focuses on the allocation of limited resources to competing demands over time with regard to costs and revenues. Because of the modeling complexity, the production planning and control system usually cannot fully capture the behavior of the base system and process and the
behavior of the customers. Hence, disturbances such as machine breakdowns, new order arrivals, order cancelations, and lack of forecast accuracy contribute to the gap between the released production plans, i.e. initial plan issued in week $t$ with production starts for weeks $t+1$ to $t+n$, and the realized plan, i.e. actually processed lots on the shop-floor. Thus, the performance of a planning algorithm and also of the corresponding planning process can be assessed by measuring the discrepancies between what has been planned and what has been realized. This measure is known as stability. A plan is called stable when the completed activities deviate as little as possible from the original plan in the face of disruptions (Pfeiffer, Kadar, and Monostori 2007; Gören and Sabuncuoglu 2008).

The proposed architecture enables interactions between the planning and simulation modules in a rolling horizon environment. Production plans are determined in a rolling horizon manner. We assume that the planning horizon is finite and shorter than the simulation horizon. Therefore, on the one side we obtain production requests for the current planning horizon after each plan determination, on the other hand we obtain completion times of lots after being processed within the simulation. Similar work is presented in Kimms (1998). A stability measure $S$ in the context of a rolling horizon setting is given by

$$S = \frac{1}{n \cdot K} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{t=1}^{\min(t_{\text{st}}, t_{\text{lt}})} \left| Q_{it}^{k} - Q_{it}^{k-1} \right| \cdot (1 - \alpha)^{t - t_{s}},$$

where $i$ is the product index, $n$ is the total number of products, $k$ is the plan index, $K$ is the total number of plans, $t$ is the period index, $t_{s}$ is the first period of plan $k$, $t_{lt}$ is the last simulated period, $L$ is the length of the planning horizon, and $Q_{it}^{k}$ is the planned quantity for product $i$ in period $t$ in plan $k$. In addition, the values of the performance measures of the plans are non-linearly weighted according to their respective time lag until the execution date. When the parameter $\alpha \in [0,1]$ is decreasing toward zero, changes close to their realization date are stronger penalized (Sridharan, Berry, and Udayabhanu 1988; Zhao, Xie, and Jiang 2001).

## 4 APPLICATION TO MASTER PLANNING APPROACHES FOR A WAFER FAB

In this section, we show how the proposed architecture can be used to assess the performance of different master planning approaches for a wafer fab. First, we introduce two production planning schemes. Then, we briefly describe the methodology for reducing the simulation model and outline some results. Afterwards, we explain the design of experiments used to generate the test instances. Finally, we present and discuss the results of computational experiments.

### 4.1 Description of Master Planning Approaches

The master planning process belongs to the mid-term planning level. It details the aggregated sales and operations plan, and it is the main input for the wafer fab scheduling and the order promising system. The master plan contains capacitated production requests for the next six months (Vieira 2006).

In this paper, we are interested in determining appropriate wafer quantities for several products and several periods of time. Note that we explicitly consider only one single wafer fab. Thus, we do not consider routing decisions since these are only relevant for manufacturing networks. The decision problem as described in Ponsignon and Mönch (2009) consists in keeping the number of unmet customer orders as low as possible and in satisfying additional forecasted demands when capacity is sufficient, while the inventory level has to be minimized. The capacity offer of the wafer fab is related to bottleneck work centers by taking re-entrant process flows of lots into account. For each bottleneck we set strict minimum and maximum loading bounds, and the sum of the processing times of lots on bottleneck machines has to be contained within this range. For simplification purposes, we assume that all products have fixed average cycle times.

We introduce two heuristic schemes for solving the master planning problem (Ponsignon and Mönch 2009). The first scheme, denoted as approach A, is a rule-based allocation algorithm. First, the product
with the highest backlog cost or the highest revenue is selected. Then, depending on the remaining capacity of the wafer fab, the demand for the selected product in the current time period is allocated. When capacity is not sufficient in the actual period, the algorithm looks for available capacity in previous time periods. This leads to pre-production and stock building but it avoids backlogs. This procedure is repeated until all products and all time periods have been considered. Note that orders are allocated with a higher priority than forecast. In addition, minimum and maximum capacity limits are strictly respected. Indeed, if the load is too low, it is increased by means of a repair loop. The second scheme, called approach B, works almost in the same way expect that products are randomly chosen without taking the costs or revenues into account. In addition, also the wafer fab loading is kept at a high level.

4.2 Reduced Simulation Models of Wafer Fabs

As shown in Section 3, the planning system is embedded in a simulation-based architecture since simulation is an effective way of representing base system and base process. However, the simulation is much more detailed than necessary for the production planning algorithms that work on aggregated data. Algorithm 1 also causes a high computational burden. In addition, mid-term planning decisions need less details than short-term production control. For these reasons, a reduced simulation model focusing on the principal characteristics of the base system and process is used in our architecture. In the following, Algorithm 2 describes the methodology proposed by Hung and Leachman (1999) and Völker and Gmilkowsky (2003) for reducing the degree of detail of simulation models.

Algorithm 2 // Methodology for reducing the simulation model

1. Determine a feasible master plan $M$
2. Transform due dates into lot release dates $M'$
3. Simulate $M'$ using a detailed simulation model $D$
4. Rank machines according to the waiting times of the lots in front of the machine and their utilization
5. Do while acceptance criteria is not met
6. Reduce $D$ to $R$ by replacing $L$ machines with low utilization or small waiting times by fixed or stochastic delays
7. Simulate $M'$ using the reduced simulation model $R$
8. Compare results from $R$ and $D$ in terms of time and solution quality.
9. If acceptance criterion is not met, then adjust $L$
10. End do
11. End algorithm 2

The methodology is applied to a variant of the MASM test data set MIMAC-I (MASM 1997). It consists of 12 machine groups, 24 distinct process flows, and 68 processing steps per process flow on average. The reduction is carried out for 120 products. The simulation horizon is 364 days.

As depicted in Algorithm 2, machine groups with long waiting times and high utilization are considered as bottlenecks. They are modeled in detail in the reduced model. All other machines are represented by fixed delays in the process flows of the lots. As a result, the reduced simulation model encompasses only four machine groups. Each process flow has an average of 29 non-delay processing steps. The reduction allows decreasing the computation time for the simulation experiments by around 30%. In order to guarantee a similar representation of the base system and process by the detailed and reduced simulation models, we compare the cycle time distribution for both models. We see from Figure 3 that both histograms are very similar. The average cycle times depicted by the red dotted lines are almost identical.

4.3 Design of Experiments

The performances of approaches A and B outlined in Section 4.1 are compared by sequentially varying the reference base system and the reference planning system in four different scenarios, respectively. For the base system, Scenarios 1 and 2 respectively refer to high and low demand levels; in Scenario 3 machine breakdowns are frequent and short, while they are rare and long in Scenario 4. For the planning sys-
tem, Scenarios 5 and 6 show cases where the capacity of the base system is over-estimated or under-estimated; in the same way Scenarios 7 and 8 correspond to situations where the cycle times that are used in the planning algorithm are over-estimated or under-estimated. We obtain eight possible factor combinations. For each of them, four independent instances of the base system and base process are created and used to assess the two approaches. The 64 simulation experiments are executed on a computer with a 2.53 GHz Intel Core2 Duo processor and 2.0 GB memory.

![Figure 3: Comparison of the cycle time distribution for detailed and reduced simulation models](image)

The planning horizon of both schemes is 26 weeks. The simulation horizon is 728 days. A new plan is determined every seven days. The experiments are carried out for 120 products in a single wafer fab using the reduced model described in Section 4.2. The standard lot size is 48 wafers. The parameter $\alpha$ in expression (1) is set to 0.6. The values of other parameters are taken from Ponsignon and Mönch (2009).

### 4.4 Computational Experiments

The performance of approaches A and B is measured by the ratio of the objective function values cumulated over the entire simulation horizon. Furthermore, the performance of both approaches is also measured by the ratio of the stability measure values given by expression (1). A ratio value higher than one refers to a decreasing planning stability by using a more sophisticated scheme rather than a simple procedure. All results are grouped according to the scenarios. The average values are shown in Table 1. It also contains additional performance measure values such as the ratio of completed wafers, the difference in percents of late wafers, and the difference in days of average cycle times as measured in the base system.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Cumul. Obj. Val.</th>
<th>Stability</th>
<th>Compl. Wafers</th>
<th>Late Wafers</th>
<th>Average CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5263</td>
<td>1.5305</td>
<td>0.9844</td>
<td>-0.30%</td>
<td>0.3 days</td>
</tr>
<tr>
<td>2</td>
<td>1.8714</td>
<td>1.0754</td>
<td>0.9691</td>
<td>0.68%</td>
<td>0.2 days</td>
</tr>
<tr>
<td>3</td>
<td>1.6118</td>
<td>1.5207</td>
<td>0.9634</td>
<td>2.23%</td>
<td>0.5 days</td>
</tr>
<tr>
<td>4</td>
<td>1.6141</td>
<td>1.5173</td>
<td>0.9629</td>
<td>2.35%</td>
<td>0.5 days</td>
</tr>
<tr>
<td>5</td>
<td>0.9563</td>
<td>1.6788</td>
<td>0.9609</td>
<td>0.59%</td>
<td>0.8 days</td>
</tr>
<tr>
<td>6</td>
<td>1.0070</td>
<td>0.7618</td>
<td>0.9818</td>
<td>2.10%</td>
<td>0.5 days</td>
</tr>
<tr>
<td>7</td>
<td>2.4012</td>
<td>1.0738</td>
<td>0.9249</td>
<td>2.62%</td>
<td>0.8 days</td>
</tr>
<tr>
<td>8</td>
<td>1.7982</td>
<td>1.0060</td>
<td>0.9876</td>
<td>4.38%</td>
<td>0.4 days</td>
</tr>
</tbody>
</table>

The ratio of the cumulated objective function values indicates that for all scenarios, except for Scenario 5, approach A outperforms approach B with respect to the objective function. The largest difference is obtained for Scenario 7 since longer cycle times increase the influence of the product selection on the objective function value. In contrast, the instances with over-estimated capacities, i.e. Scenario 5, offer some advantage for the scheme that tends to produce more than necessary for later demand fulfillments.
Moreover, the ratio of the stability measures shows that in most cases approach B produces more stable plans in the course of time than approach A. It is particularly true when the simple allocation scheme is able to build stocks in advance for a wide variety of products to deal with forthcoming disruptions of customer demands. In addition, approach A tends to always concentrate on the same group of products in the long-term because of the selection rule, while it is forced to make large changes in the short-term to react to new demands and backorders. Note that in Scenario 6 the reduced capacity offer gives fewer opportunities to approach B for pre-production, which leads to a poor stability performance.

Furthermore, the ratio of completed wafers points out that approach B proposes to produce more finished goods that are stored. It reduces the objective function value because of inventory holding costs, but it increases the reactivity. Indeed, there are less late wafers and the average cycle times are shorter.

5 CONCLUSION AND OUTLOOK FOR FUTURE RESEARCH

We presented an architecture for simulation-based performance assessment of production planning approaches. The core of the proposed architecture is a blackboard-like data layer that is used as an interface between the simulation model and the production planning and control systems. The algorithm associated with the architecture was also outlined. The planning performance is evaluated by means of a non-linearly weighted stability measure in a rolling horizon environment. We introduced two master planning approaches for a single semiconductor wafer fab. The first scheme is a rule-based allocation procedure, while in the second scheme products are randomly chosen and pre-production is favored. Moreover, a reduced simulation model focusing on the principal characteristics of the base system and process is used within the architecture. The methodology used to decrease the level of detail of the simulation model without altering the quality of the results was briefly presented. Computational results were also provided for the reduced model. Finally, the performances of both planning approaches were assessed with the help of the proposed architecture for eight different scenarios. From the results of computational experiments we concluded that the first scheme reaches larger objective function values, while the second scheme produces more stable plans in most cases.

There are some directions for future research. We intent on expanding the base system to a network of wafer fabs and to incorporate the routing decision in the planning system. Then, we would like to investigate the possibility of using the proposed architecture to obtain more realistic cycle times by iteratively alternate between planning and simulation. Finally, it is interesting to consider the influence of stochastically defined forecast on master planning approaches in a rolling horizon environment.

REFERENCES

Ponsignon and Mönch


AUTHOR BIOGRAPHIES

THOMAS PONSIGNON is a Ph.D. candidate in the Department of Mathematics and Computer Science at the University of Hagen, Germany. He is also working as a part-time employee at Infineon Technologies AG in the field of supply chain management. He received master’s degrees in industrial engineering from the EPF-Ecole d’Ingénieurs, Sceaux, France and the University of Applied Sciences, Munich, Germany. His research interests include production planning and simulation for semiconductor manufacturing networks. His email address is <Thomas.Ponsignon@infineon.com>.

LARS MÖNCH is a Professor in the Department of Mathematics and Computer Science at the University of Hagen, Germany. He received a master’s degree in applied mathematics and a Ph.D. in the same subject from the University of Göttingen, Germany. His current research interests are in simulation-based production control of semiconductor wafer fabrication facilities, applied optimization, and artificial intelligence applications in manufacturing, logistics, and service operations. He is a member of GI (German Chapter of the ACM), GOR (German Operations Research Society), SCS, INFORMS, and IIE. His email address is <Lars.Moench@fernuni-hagen.de>.