BORDERLESS FAB SCENARIOS IN HIERARCHICAL PLANNING SETTINGS: A SIMULATION STUDY

Raphael Herding[1,2], and Lars Mönch[1,3]

[1]Forschungsinstitut für Telekommunikation and Kooperation (FTK), Dortmund, GERMANY
[2]Westfälische Hochschule, Bocholt, GERMANY
[3]Dept. of Mathematics and Computer Science, University of Hagen, Hagen, GERMANY

ABSTRACT

Lots are transferred in borderless fab (BF) scenarios from one wafer fab to another nearby fab to carry out process steps of the transferred lots. BF aims to compensate for scarce bottleneck capacity in some of the wafer fabs. One goal of the master planning function is to distribute the demand over the wafer fabs such that situations are avoided where large queues of lots arise in the wafer fabs. Due to inaccurate modeling of capacities and lead times in master planning, this goal is not always reached. Wafer fabs are often heretogenous. This leads to additional costs for BF scenarios which might make them less attractive. We are interested in exploring conditions with respect to master planning and wafer fab heterogeneity under which BF scenarios are still beneficial. Master planning, production planning, and the BF lot transfers are carried out in a rolling horizon setting using a cloud-based infrastructure.

1 INTRODUCTION

Master planning deals with determining which quantity of a certain product must be completed in which front-end or back-end nodes of a semiconductor network (Mönch et al. 2018b) while production planning deals with determining releases into each single wafer fab such that demand is met or instructions of higher planning levels are fulfilled and some performance measure of interest such as profit or cost is optimized (Missbauer and Uzsoy 2020). The finite capacity of each wafer fab and the long cycle time of the products are major constraints in the hierarchical planning process including master and production planning. In semiconductor manufacturing, cycle times are defined as the time span between releasing work into wafer fabs and its emergence as final products.

Production planning activities are performed for each single wafer fab of the network. However, when wafer fabs are located geographically close to each other there are settings possible where specific process steps of some lots can be performed in one of the neighboring wafer fabs. Lots are the moving entities in wafer fabs. Each lot consists of a given set of wafers, thin silicon discs on which integrated circuits are built layer by layer. Such settings, known as BF scenarios in the literature (Mönch et al. 2013, Mönch et al. 2018a), can be found in semiconductor supply chains in Asia and Europe. However, despite their importance, BF scenarios are only rarely studied. Moreover, often simplifying assumptions are made (cf., Lendermann et al. 2004; Gan et al. 2007; Herding and Mönch 2023). For instance, typical assumption are that only two identical wafer fabs are considered and that the interaction of the BF activities with higher-level planning activities are not considered.

In the present paper, we extend the multi-agent system (MAS) approach from Herding and Mönch (2023) for BF scenarios with two identical wafer fabs and production planning towards a hierarchical setting where master planning and heterogeneous wafer fabs are included. We are interested in exploring the degree of heterogeneity of the involved wafer fabs and the role of master planning with respect to the performance advantages which can be obtained from allowing BF activities.

The paper is organized as follows. The problem at hand is described in the next section. This includes a discussion of relevant work. In Section 3, we present the hierarchical approach and the extensions of the

979-8-3315-3420-2/24/$31.00 ©2024

MAS. The results of simulation experiments applying the planning approaches in a rolling horizon setting are discussed in Section 4. Finally, conclusions are future research directions are presented in Section 5.

2 PROBLEM SETTING

2.1 Problem

We assume that $k = 1, \dots, m$ not necessarily identical wafer fabs can participate in the BF setting. We distinguish delivering wafer fabs from consuming ones. Delivering wafer fabs have heavily overloaded bottleneck work centers whereas the consuming wafer fabs have no overload at their work centers. Here, a work center is a group of machines that offer the same functionality. A wafer fab cannot be a delivering and a consuming fab at the same time. If the number of lots in front of the bottleneck of the delivering wafer fab k n_k exceeds a threshold Δ_k at a certain point in time then $n_k - \Delta_k$ lots are transferred from the bottleneck work center of wafer fab k to an appropriate work center of the consuming fab l. The exchanged lots are the ones with the smallest local due dates in k. These lots are then processed by the machines of the target work center of wafer fab l. After processing in wafer fab l, the lots are transferred back to wafer fab k for further processing. The exchange can be repeated if the bottleneck work center is visited several times by the same lot, i.e. to deal with the reentrant flows which exist in all wafer fabs (Mönch et al. 2013).

First, we discuss a setting where the participating wafer fabs belong to a single company. We refer to this as intra-company setting. Master planning determines which quantity of a certain product must be finished in which period in the front-end or back-end nodes of a semiconductor network (Mönch et al. 2018b). Master planning interacts with production planning since the output of the master planning function is used as input for the production planning activities in each single node of the network. Perfect master planning decisions would result in a situation where overloads of wafer fabs are unlikely. But since master planning is based on a rough modeling of the available capacity of the nodes, on eventually erroneous lead times, and demand uncertainty, the master planning decisions might lead to situations where overload situations occur in certain wafer fabs.

Second, wafer fabs participating in a BF setting can belong to different companies. In this situation, there are no joint master planning activities for the wafer fabs. We refer to this as inter-company setting. Of course, it is also likely that the wafer fabs are heterogeneous in this setting since different companies are involved.

A BF setting might be helpful to mitigate the consequences of the overloads in both situations. Since we know from Herding and Mönch (2023) that it is beneficial if production planning takes into account the BF activities the following two approaches are investigated in the present paper:

1. **Reference scenario with no borderless fab (N-BF):** There is no lot transfer between the wafer fabs. Production planning will be carried out for each of the participating wafer fabs. Production planning is used to adjust the overload situations found in the wafer fabs.
2. **Borderless fab scenario with advanced production planning (BF-APP):** Lots are exchanged between delivering and consuming wafer fabs. This exchange is also considered when the production planning models of the wafer fabs are generated. The available capacity in the first period of the planning window will be correctly modeled in the two production planning models.

Therefore, the research questions investigated in this paper are as follows:

1. **Intra-company setting:** Under which levels of erroneous master planning decisions and wafer fab heterogeneity a BF setting will lead to performance improvements, i.e. larger network-wide profit?
2. **Inter-company setting:** Under which levels of wafer fab heterogeneity a BF setting will lead to larger network-wide profit?

To answer these questions, we extend the MAS proposed by Herding and Mönch (2023) by introducing a master planning decision-making agent and a corresponding staff agent. Moreover, we will conduct simulation experiments with the MAS to assess the performance of the interaction of network-wide master planning, fab-specific production planning, and the application of the BF setting to mitigate overload situations in the participating wafer fabs.

2.2 Discussion of Related Work

We discuss related work with respect to BF settings and hierarchical planning approaches for semiconductor supply chains. Lendermann et al. (2004) and Gan et al. (2007) analyze BF scenarios using distributed simulation. The consequences of different lot batching sizes for the cross-fab process step on lot transfer frequency and cycle time are investigated. However, planning is not considered. Only homogenous wafer fabs are taken into account.

Wu and Chen (2007) and Wu and Chen (2008) discuss an approach that exchanges capacity among several wafer fabs that are within close geographical proximity. A simulation-based trading method for two wafer fabs is designed that allow for capacity sharing of certain work centers. A game theory-based approach is presented by Chien and Kuo (2013) for a similar problem. However, master and production planning activities are not considered in these papers. Although heterogeneous wafer fabs are assumed in these papers, the impact of the degree of heterogeneity is not investigated.

Hierarchical planning approaches are only rarely discussed in the literature for semiconductor supply chains (Mönch et al. 2018b). We are only aware of Herding and Mönch (2022) where the interaction of master planning and production planning is studied using a MAS-based infrastructure and Herding and Mönch (2024) where the interaction of master planning and demand fulfillment is investigated. In the present paper, we reuse the master planning and production planning implementations from Herding and Mönch (2022). However, we must add the BF activities.

The most pertinent work for the present paper is Herding and Mönch (2023). Here, the two present authors describe a MAS for a BF setting with two identical wafer fabs and an artificially created overload situation in one of them. In the present paper, we extend the MAS by adding software agents for master planning. Moreover, the BF scenarios include heterogeneous wafer fabs.

3 PLANNING APPROACHES WITH BF ACTIVITIES

3.1 Overall Approach and MAS Extensions

Both intra- and the inter-company settings are considered in this paper. The main difference between the two settings is how the overload in the delivering wafer is created. In the inter-company setting, the demand for overload is synthetically created as already specified in Herding and Mönch (2023). The goal of investigating this setting in the present paper again is to repeat the simulation experiments for heterogeneous wafer fabs to determine to which extent the heterogeneity reduces the improvements. This setting is shown in Figure 1. The dashed lines between the different wafer fabs indicate that these wafer fabs belong to different companies. Note that wafer fab 2 is the consuming one in Figure 1, whereas the wafer fabs 1 and 3 are the delivering ones.

In the intra-company setting, the overload is caused by inappropriate master planning decisions. The master planning decisions result in the quantities of the different products and periods for the production planning function to be taken into account for the different wafer fabs. In a certain sense, wafer fab-specific desired output quantities are determined. BF activities can be used to mitigate the impact of erroneous master planning decisions. The overall situation is shown in Figure 2. Again, wafer fab 2 is the consuming one, whereas the wafer fabs 1 and 3 are the delivering ones.

The master planning decision-making agent (DMA) is responsible for determining a master plan. It coordinates the different fab agents as well as the mid-term network-wide planning agent.

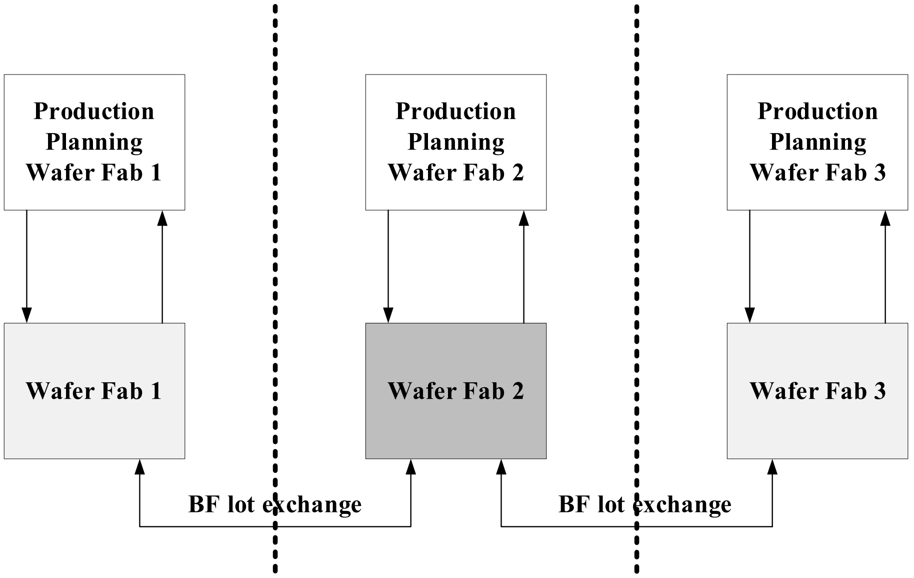
Production Planning Wafer Fab 1
Production Planning Wafer Fab 2
Production Planning Wafer Fab 3
Wafer Fab 1
Wafer Fab 2
Wafer Fab 3
BF lot exchange
BF lot exchange

Figure 1: Inter-company setting with BF activities.

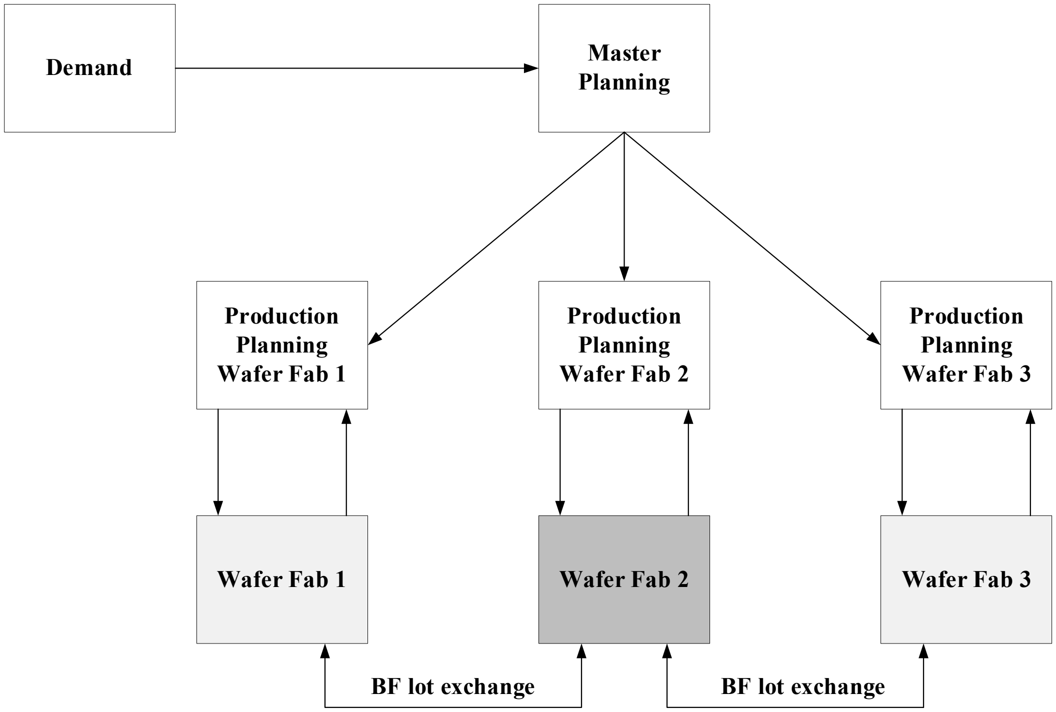
Demand
Master Planning
Production Planning Wafer Fab 1
Production Planning Wafer Fab 2
Production Planning Wafer Fab 3
Wafer Fab 1
Wafer Fab 2
Wafer Fab 3
BF lot exchange
BF lot exchange

Figure 2: Intra-company setting with BF activities.

The fab agents are DMAs that make production planning decisions for the different wafer fabs, while the mid-term network-wide planning agent is a staff agent (SA). This software agent provides the desired output quantities per product and period to the corresponding fab agents. The mid-term network-wide planning agent supports the master planning DMA. It prepares to perform the planning algorithm, it runs the planning algorithm itself, and provides the computed master plans to the master planning DMA. This is shown in Figure 3 by means of an Unified Modeling Language (UML) sequence diagram. The different activities will be described in more detail in the following subsections.

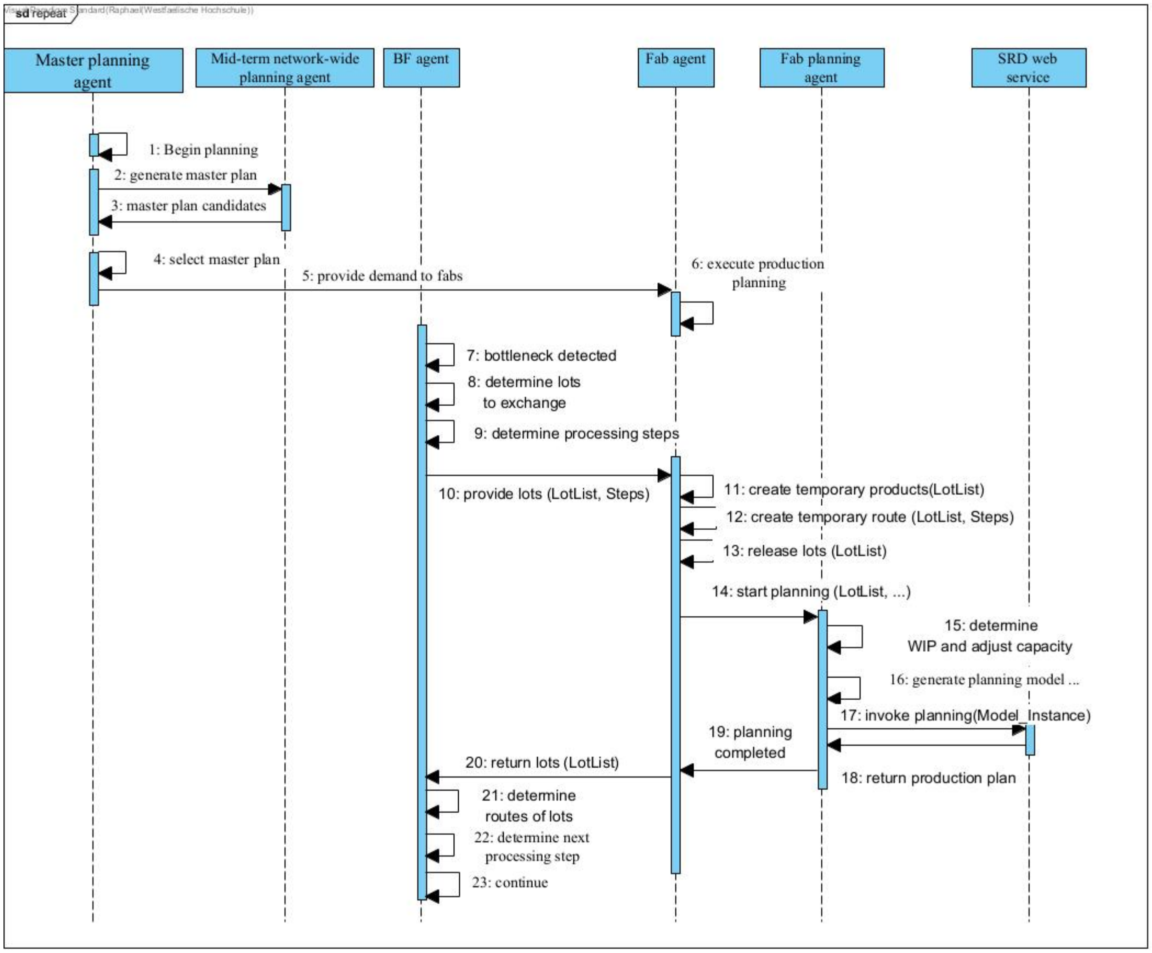
sd repeat
Master planning agent
Mid-term network-wide planning agent
BF agent
Fab agent
Fab planning agent
SRD web service
1: Begin planning
2: generate master plan
3: master plan candidates
4: select master plan
5: provide demand to fabs
6: execute production planning
7: bottleneck detected
8: determine lots to exchange
9: determine processing steps
10: provide lots (LotList, Steps)
11: create temporary products(LotList)
12: create temporary route (LotList, Steps)
13: release lots (LotList)
14: start planning (LotList, ...)
15: determine WIP and adjust capacity
16: generate planning model ...
17: invoke planning(Model_Instance)
18: return production plan
19: planning completed
20: return lots (LotList)
21: determine routes of lots
22: determine next processing step
23: continue

Figure 3: Sequence diagram for the interaction of the different planning agents and the BF-related activities.

3.2 Modeling of the BF Scheme

The BF agent is responsible for detecting overload situations in the delivering wafer fab. The work center agent of the bottleneck work center of this wafer fab continuously informs the BF agent about the queue length and the utilization of the machines of the work center. Whenever an overload situations is observed, the BF agent decides which lots have to be delivered to the consuming wafer fab. The process steps that must be performed in the consuming wafer fab are then determined. Moreover, the fab agent of the consuming wafer fabs has to be informed about the exchange, and the data of the affected lots is then sent to the fab agent. This agent immediately launches the received lots into the base system of the consuming wafer fab. The BF agent keeps track of the lots belonging to the delivering wafer fab that are processed in a consuming wafer fab. The base system of the consuming wafer fab is frequently updated because lots are are no longer necessary if the lots are completed and sent back to the delivering wafer fab.

3.3 Production Planning

In the following, we assume that a production planning window of length $T^{(PP)}$ with equidistant periods labeled by $t = 1, \dots, T^{(PP)}$ exists. We apply the simple rounding down (SRD) planning formulation proposed by Kacar et al. (2013) as a production planning approach in the different fab planning agents. The approach is based on assuming fixed exogenous lead times (LTs) that are an integer multiple of the period length. An objective function that considers work in progress (WIP), backlog, and inventory cost is used. Capacity constraints must be fulfilled for each work center in a wafer fab. For the details of the linear programming (LP) formulation in the BF context we refer to Herding and Mönch (2023).

The fab planning agent of a consuming wafer fab receives the message from its corresponding fab agent and interprets the messages regarding the transferred lots. An actual instance of the SRD planning model is generated by the fab planning agent. The instance must be enriched by the received data, i.e., temporary routes, process steps, and products are created in the base system of the consuming wafer fab for the transferred lots from a delivering wafer fab. This enrichment process results in changes of the capacity usage which appears as initial WIP in the planning instance since the received lots are already released. After the generation of the LP model, the instance is transferred to the web service in the MAS which is responsible for solving the LP. The computed production plan is sent back to the fab planning agent.

When a new production plan is required in a delivering wafer fab between Steps 10 and 20 of Figure 3, the fab planning agent of the delivering wafer fab requests capacity information of the transferred lots by the BF agent. This capacity information is important to estimate at which point in time the transferred lots will be sent back to a delivering wafer fab since more capacity is required to continue processing the previously transferred lots within the Steps 21-23.

3.4 Master Planning

In the remainder of this paper, we assume that a master planning window of length $T^{(MP)}$ with equidistant periods labeled by $t = 1, \dots, T^{(MP)}$ exists. We consider only a set of wafer fabs that are working in parallel, back-end facilities are neglected. Moreover, for the sake of simplicity, we assume that the period length is the same for both master planning and production planning. We apply the master planning formulation described by Herding and Mönch (2021). It is based on the assumption that only the finite capacity of the bottleneck work centers of the nodes of the semiconductor supply chain is considered in the resulting LP models. Moreover, again fixed exogenous LTs are assumed that are an integer multiple of the period length. As a result, the offered capacity of the bottleneck work center that forms the right-hand side in the capacity constraints might be over- or underestimated. Furthermore, we consider only one demand class for the sake of simplicity that is the joint demand. Moreover, the LT information for some of the products might be wrong. Instead of assuming individual demand for the wafer fabs as in the inter-company setting, joint demand is given for all the parallel wafer fabs in the intra-company setting. Based on this joint demand, master planning computes the desired output quantities per product and period for the different wafer fabs. We use a profit-based objective function that is given by the differences of revenue, backlog, inventory, and production cost. Due to space limitation, we do not recall the master planning formulation. Instead of this, we refer to Herding and Mönch (2021) for the details of the formulation. It is straightforward to change this formulation in such a way that only nodes for wafer fabs are taken into account.

4 SIMULATION EXPERIMENTS

4.1 Simulation Models

To consider more real-world like BF scenarios, three different wafer fab simulation models are applied. They are combined into a single simulation model for the set of parallel wafer fabs. The first submodel of this model is a slightly simplified version of the wafer fab part of the semiconductor supply chain simulation testbed proposed by Ewen et al. (2017) that is publicly available under Simulation Models (2024). The model is similar to the MIMAC Data Set 1 (cf. Fowler and Robinson 1995). It consists of a wafer fab with more than 200 machines. Two products with more than 200 process steps are processed on machines that are organized in around 80 work centers. The model contains batch processing machines and highly reentrant process flows. A batch is a group of lots that are processed at the same time on the same machine. Exponentially distributed machine breakdowns are taken into account. The cycle times of the two products are between two and three weeks depending on the utilization of the planned bottleneck work center which is formed by the steppers. First-in-first-out (FIFO) dispatching is used at all work centers.

The second submodel is a reduced variant of the MIMAC Data Set 1 (cf. Simulation Models 2024) that contains two routes with 103 and 100 steps, respectively. The process flow is highly reentrant. The jobs are processed on 140 machines that are organized into 48 work centers. There are batch processing machines among the machines found in the model. The cycle time of the two products is between seven until nine days depending on the utilization of the planned bottleneck which is again the stepper work center. FIFO dispatching is used at all work centers. We refer to this model as Half-Fab, since it contains around one half of the processing steps of the MIMAC 1 simulation model.

The third sub model is a multi-product version of the first submodel. It has 32 products, whereas the machinery is the same as for the first submodel. The process flows of the two products are divided into subflows. Subflows are randomly chosen from the two products until 32 products are obtained. We refer to Mönch and Zimmermann (2011) for the details of the product generation procedure. The characteristics of the three submodels are summarized in Table 1. The number of layers indicates how often the stepper work center is visited, the planned bottleneck in all the simulation submodels. The commercial simulation engine AutoSched AP is applied in the simulation experiments.

Table 1: Main characteristics of the applied simulation submodels.

Submodel	#Products	#Work centers	#Machines	#Layers
Semiconductor Supply Chain Data Set	2	80	200	8-11
Half-Fab	2	48	140	5-6
Multi-product MIMAC 1	32	80	200	6-8

4.2 Design of Experiments

In this research, we are interested in investigating intra- and inter-company scenarios. The former scenario is used to evaluate the use of a BF setting considers all the three described wafer fabs of the simulation model. The wafer fabs are of course heterogeneous. We over- or underestimate the capacity as well as the LT in order to investigate possible applications of BF scenarios. In the first scenario, the first submodel, i.e. the model from the semiconductor supply chain data set, is the consuming wafer fab where the other two wafer fabs are the delivering wafer fabs. We use $\Delta_2=18$ and $\Delta_3=32$ lots as threshold values to transfer lots from the wafer fabs associated with the second and third simulation submodel to the first submodel.

The inter-company setting investigates the case where no central master planning is available. The output targets for the different wafer fabs are independent for each involved company. In this scenario, we use only two wafer fabs. We take the first and third submodel of the joint simulation model, where the third submodel is used as delivering wafer fab, and the first submodel serves as consuming wafer fab. We apply again $\Delta_3=32$.

In a realistic scenario, transferring and processing lots in different wafer fabs cause costs of different types. That means that a BF scenario should not be treated as a pure capacity extension. For instance, the setup time of a machine will increase when a lot of another wafer fab will be processed in this wafer fab.

This increase in setup time is caused by configuration and/or qualification activities on the work centers in the consuming fab where the processing will take place. To model this behavior, we consider an additional setup time for each lot transfer. The setup time only occurs in the consuming wafer fab since the lots have to be transferred to and processed there. The setup time is defined as

$$s(k) \coloneqq a + b/2^k,$$

where a is a base setup time that occurs only for the first transfer of a specific product, and b is a time-dependent setup time part. The second parameter b is used to model the effect that it may not need the same time to setup a machine when a lot of the same delivering wafer fab has been processed on the machine in the past, i.e., we take into account learning effects in a simple manner. Both parameters are given in machine hours. The k quantity is defined as the number of times a lot of a certain product of the same delivering fab has been processed on that machine.

The lot transfer scheme and the production planning models are carried out in a rolling horizon setting using discrete-event simulation applying the cloud-based infrastructure proposed by Herding and Mönch (2022). A simulation horizon of a single year is applied together with a planning window that consists of 26 periods each of them with a length of a single day. We use initial WIP taken from long simulation runs to initialize the simulation.

Normally distributed stationary demand for the entire network is used that results in 93% - 96% planned bottleneck utilization (BNU) in the delivering wafer fabs. The coefficient of variation (CV) of the demand is 0.25. A product mix of 1:1 is used. The consuming wafer fab has a planned BNU of around 85% to 90%. This BNU level is reached by underestimating the bottleneck capacity in this wafer fab. Local due dates of the lots are set using a forward termination scheme with a prescribed flow factor value (Mönch et al. 2013) that is appropriate for the reached BNU levels.

We are interested in maximizing the profit, i.e. the difference of revenue and the sum of WIP, inventory, and backlog costs on the network level. We apply the unit backlog cost b_{gt}:=50 for product g in period t, unit WIP cost ω_{gt}: =20, and unit inventory holding cost $h_{gt}: = 15$ for production planning in the simulation experiments. Moreover, the revenue per lot is $r_{gt} \coloneqq 180$. For master planning, we apply for a unit of revenue $r_{1gt}: = 180$ where the 1 indicates the single demand class, a unit holding cost $h_{gt}: = 15$, unit production cost $m_{gj}: = 20$ for wafer fab j, and unit backlog cost $ud_{1gt}: = 50$. As the additional setup time in hours, we chose

$$s(k) \coloneqq \begin{cases} 4, & \text{for } k = 0, \\ 1/2^k, & \text{for } k > 0. \end{cases}$$

for the kth lot transfer of a certain product.

Five independent demand instances are used in the simulation experiments. Moreover, ten independent simulation replications are performed for each demand instance to compute the performance measure values in the face of execution uncertainty. The average profit is taken over all replications and all simulation submodels.

We are interested in assessing the performance of the two described scenarios. We compare each of the results to the case when no BF is used. The design of experiments for the first scenario is summarized in Table 2.

Table 2: Design of experiments for the intra-company scenario.

Factor	Level	Count
BNU	high	1
CV of demand	0.25	1
Capacity setting	overestimated (OE), underestimated (UE), correct (C)	3
Lead time setting	overestimated (OE), underestimated (UE), correct (C)	3
Independent demand		5
Independent simulation		10
Total number of simulation runs		450

We are interested in investigating the effect of over- or underestimating the capacity and the lead time inside the master planning function. We use 75% of the available capacity in the case of underestimated capacity. The same number, i.e. 75%, is used when the LT is underestimated in the master planning formulation. In the case of overestimating the capacity, we use 125% of the available capacity within master planning. The same number is again used for overestimating the LT in the master planning formulation.

On the one hand, we expect from scenario 1 that the BF setting increases the profit in situations where an overload of the delivering fabs exists. On the other hand, we expect that a BF setting is not useful in situations where long queues in front of a work center do not exist.

The design of experiments for the second scenario is much simpler. Since we do not consider master planning, we are only interested in investigating the effect of a BF setting when we introduce the additional setup time in the consuming wafer fab. We compare it against the non-BF setting. Normally distributed stationary demand is used that results in 93% - 96% planned bottleneck utilization (BNU) in the delivering wafer fab. The consuming wafer fab has a planned BNU of around 85% to 90%. Moreover, we use $\Delta_2=32$ lots as threshold value. All the other parameter values are similar to scenario 1. This is especially true for the parameter values a and b of the additional setup time.

Since we do not consider master planning, we report the profit obtained by decisions of the production planning function. Again, we expect that the BF scenario will result in higher profit if a bottleneck situation occurs and at the same time the setup time increase is not too high.

4.3 Simulation Results

We observe from the simulation results that the advantage of using a BF setting depends on various factors. Table 3 shows the results for the first scenario. The values are relative to the non-BF scenario. 95% confidence intervals are presented instead of the values of point estimates to obtain statistically reasonable results.

In the case of the results that are marked bold, a BF setting is beneficial. Non-bold marked results show that a BF setting is not beneficial. Overall, a BF setting is useful when the delivering wafer fabs are overloaded. When the capacity is overestimated (OS) and the lead time is underestimated (US) a BF setting leads to the highest profit increase (around 23%) in comparison to all other scenarios. This is reasonable since the utilization of the delivering wafer fabs is the highest due to scare capacity and underestimated LTs. In another case where the capacity is overestimated and the LT is set in a correct manner, a BF setting also leads to some improvement. In situations where either the capacity is overestimated and the LT is correct or the LT is underestimated and the capacity is correct, a BF setting is beneficial. This is reasonable since both situations lead to long queues in front of the bottleneck work center of the delivering wafer fabs.

In situations where no overload exists, a BF setting is not beneficial. When the capacity is underestimated and the LT is overestimated at the same time, it is very likely that an overload does not exist. If there is no bottleneck, a BF setting is not beneficial at all.

Table 3: Simulation results for the for the intra-company scenario.

	LT		
Capacity	OS	US	C
OS	**0.874 ± 0.037**	**0.770 ± 0.044**	**0.808 ± 0.05**
US	0.976 ± 0.081	0.933 ± 0.029	0.979 ± 0.042
C	0.953 ± 0.065	**0.842 ± 0.080**	1.022 ± 0.085

The second scenario shows a similar behavior as the setting with identical wafer fabs working in parallel (cf. Herding and Mönch 2023). It leads to a profit increase of up to 8.71% compared to the non-BF setting. As expected this value is smaller than the improvement of 11.3% obtained for the BF scenario with identical wafer fab investigated by Herding and Mönch (2023).

5 CONCLUSIONS AND FUTURE RESEARCH

In this paper, we analyzed intra- and inter-company BF scenarios for multiple heterogeneous wafer fabs. An existing MAS from previous research was extended towards hierarchical planning using master planning and production planning. In addition, the lot exchange using a hierarchically organized MAS was described.

We demonstrated by simulation experiments applying the planning approaches and the BF approaches in a rolling horizon manner that it is worth to exchange lots between wafer fabs when the delivering fabs are overloaded even if an exchange of lots is penalized in a heterogeneous wafer fab situation. We observed that both the intra- and inter-company BF setting lead to higher profit relative to a setting where lots are not exchanged across wafer fabs.

There are several directions of future research. First of all, we believe that it is desirable to fully automate the generation of the resulting LP models for both master and production planning using an appropriate ontology for semiconductor supply chain planning. At the same time, it is also desirable to make decisions on the lot exchange itself in a more dynamic way, i.e. directly in the planning formulations or related scheduling models, rather than using the myopic rule-based approach applied in the present paper. Learning effects can be considered in the LP models for production planning similar to Ziarnetzky and Mönch (2016). It is also interesting to design negotiation approaches, for instance, for sharing capacity among the different wafer fabs using the proposed MAS prototype.

ACKNOWLEDGMENTS

The authors would like to thank John Fowler, Arizona State University, for fruitful discussions on BF settings in semiconductor manufacturing. They were supported by the SC4EU project which receives funding from the CHIPS JU under grant agreement No. 101139949.

REFERENCES

Chien, C.-F. and R.-T. Kuo. 2013. “Beyond Make-or-buy: Cross-company Short-term Capacity Backup in Semiconductor Industry Ecosystem”. *Flexible Services and Manufacturing Journal* 25(3): 310–342.

Ewen, H., L Mönch, H. Ehm, T. Ponsignon, J. W. Fowler, and L. Forstner. 2017. “A Testbed for Simulating Semiconductor Supply Chains”. *IEEE Transactions on Semiconductor Manufacturing* 30(3): 293-305.

Fowler, J. W. and J. Robinson. 1995. “Measurement and Improvement of Manufacturing Capacity (MIMAC) Final Report”. Technology Transfer #95062861A-TR, SEMATECH.

Gan, B., M. Liow, A. Gupta, P. Lendermann, S. Turner, and X. Wang. 2007. "Analysis of a Borderless Fab Using Interoperating AutoSched AP Models". *International Journal of Production Research* 45(3): 675-697.

Herding, R. and L. Mönch. 2021. A Short-Term Demand Supply Matching Approach for Semiconductor Supply Chains. INFORMATIK BERICHTE 382 – 06/2021, University of Hagen, Department of Mathematics and Computer Science. https://www.fernuni-hagen.de/imperia/md/content/fakultaetfuermathematikundinformatik/forschung/berichte/bericht_382.-pdf, accessed May, 2nd 2024.

Herding, R. and L. Mönch. 2022. "An Agent-based Infrastructure for Assessing the Performance of Planning Approaches for Semiconductor Supply Chains". *Expert Systems with Applications* 202: 117001.

Herding, R. and L. Mönch. 2023. "Agent-Based Decision Support in Borderless Fab Scenarios in Semiconductor Manufacturing". In *2023 Winter Simulation Conference (WSC)*, 2230-2241 https://doi.org/10.1109/WSC60868.2023.10407842.

Herding, R. and L. Mönch. 2024. "A Rolling Horizon Planning Approach for Short-Term Demand Supply Matching". *Central European Journal of Operations Research* 32: 865–896.

Lendermann, P., B.-P. Gan, Y. L. Loh, T. Sip, K. Lieu, J. W. Fowler, and L. F. McGinnis. 2004. "Analysis of a Borderless Fab Scenario in a Distributed Simulation Testbed". In *2004 Winter Simulation Conference (WSC)*, 1896-1901 https://doi.org/10.1109/WSC.2004.1371546.

Kacar, N. B., L. Mönch, and R. Uzsoy. 2013. "Planning Wafer Starts using Nonlinear Clearing Functions: a Large-Scale Experiment". *IEEE Transactions on Semiconductor Manufacturing* 26(4):602-612.

Missbauer, H. and R. Uzsoy. 2020. *Production Planning with Capacitated Resources and Congestion*. New York: Springer.

Mönch, L., J. W. Fowler, and S. J. Mason. 2013. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*. New York: Springer.

Mönch, L., R. Uzsoy, and J. W. Fowler. 2018a. "A Survey of Semiconductor Supply Chain Models Part I: Semiconductor Supply Chains and Strategic Network Design". *International Journal of Production Research* 56(13):4524-4545.

Mönch, L., R. Uzsoy, and J. W. Fowler. 2018b. "A Survey of Semiconductor Supply Chain Models Part III: Master Planning, Production Planning, and Demand Fulfillment". *International Journal of Production Research* 56(13):4524-4545.

Mönch, L. and J. Zimmermann. 2011. "A Computational Study of a Shifting Bottleneck Heuristic for Multi-Product Complex Job Shops". *Production Planning & Control* 22(1): 25-40.

Simulation Models. 2024. https://p2schedgen.fernuni-hagen.de/downloads/simulation, accessed May, 2nd 2024.

Wu, M.-C. and W.-J. Chang 2007. "A Short-Term Capacity Trading Method for Semiconductor Fabs with Partnership". *Expert Systems with Applications* 33: 476-483.

Wu, M.-C. and W.-J. Chang. 2008. "A Multiple Criteria Decision for Trading Capacity Between Two Semiconductor Fabs". *Expert Systems with Applications* 35: 938–945.

Ziarnetzky, T. and L. Mönch, L. 2016. "Incorporating Engineering Process Improvement Activities into Production Planning Formulations Using a Large-Scale Wafer Fab Model". *International Journal of Production Research* 54(21): 6416-6435.

AUTHOR BIOGRAPHIES

RAPHAEL HERDING is a Professor for Software Engineering at the Westfälische Hochschule Bocholt. He received a master's degree in applied computer science and a Ph.D. in computer science from the University of Hagen, Germany. His current research interests are in multi-agent systems, cloud computing, and supply chain management, especially for the semiconductor industry. His email address is raphael.herding@w-hs.de.

LARS MÖNCH is full professor of Computer Science at the Department of Mathematics and Computer Science, University of Hagen where he heads the Chair of Enterprise-wide Software Systems. He holds M.S. and Ph.D. degrees in Mathematics from the University of Göttingen, Germany. After his Ph.D., he obtained a habilitation degree in Information Systems from Technical University of Ilmenau, Germany. His research and teaching interests are in information systems for production and logistics, simulation, scheduling, and production planning. His email address is Lars.Moench@fernuni-hagen.de. His website is https://www.fernuni-hagen.de/ess/team/lars.moench.shtml.