Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

AGENT-BASED DECISION SUPPORT IN BORDERLESS FAB SCENARIOS IN SEMICONDUCTOR MANUFACTURING

Raphael Herding

Forschungsinstitut für Telekommunikation
und Kooperation e. V., Westfälische Hochschule
Wandweg 3
Dortmund, 44149, GERMANY

Lars Mönch

Forschungsinstitut für Telekommunikation
und Kooperation e. V., University of Hagen
Wandweg 3
Dortmund, 44149, GERMANY

ABSTRACT

The design and the implementation of a multi-agent system (MAS) for a borderless fab scenario is described. In such a scenario, lots are transferred from one wafer fab to a nearby one to perform process steps of the transferred lots. Production planning is carried out individually for each of the wafer fabs. The modeling of the available and requested capacity in the production planning models of the participating wafer fabs is affected by the lot transfer. The transfer of the route information from one wafer fab to another to automatically generate the linear programming models is described. Production planning is carried out in a rolling horizon setting using a cloud-based infrastructure. We show by simulation experiments with the MAS with a correct modeling of the capacity in production planning results in improved profit compared to a setting where the lot transfer is not taken into account in the planning formulations.

1 INTRODUCTION

Production planning deals with determining releases into a wafer fab such that demand is met and some performance measure of interest such as profit or cost is optimized (Missbauer and Uzsoy 2020; Mönch et al. 2018a). The finite capacity of the wafer fabs and the long cycle time of the products must be taken into account as major constraints. Here, the cycle time is the time span between releasing work into a wafer fab and its emergence as final product. Release planning is an important planning function in semiconductor supply chains (Mönch et al. 2018b). Production planning takes place for each individual wafer fab. However, when wafer fabs are located close to each other it is possible that specific process steps of some lots can be performed in the neighboring wafer fabs. Such settings can be found quite often in real-world semiconductor supply chains, for instance in Singapore, Taiwan, or Germany. These setting are called borderless fab scenarios in the literature (Mönch et al. 2013). However, to the best of our knowledge despite their importance production planning and control problems for borderless fab scenarios are only rarely studied in the literature. We are only aware of Lendermann et al. (2004) and Gan et al. (2007) where the impact of different lot batching sizes for the cross-fab process step on lot transfer frequency and cycle time is studied.

A single wafer fab can be modeled as a complex job shop consisting of a large number of different work centers. A single work center consists of machines that offer the same functionality, i.e., these machines are parallel machines. Silicon wafers, thin discs of 200 or 300 mm diameter, are used to produce several thousands of chips in a layer-by-layer manner. Up to 800 process steps are required on the machines of the work centers to manufacture a wafer. Lots, a group of up to 50 wafers, are the moving entities in wafer fabs. In the present paper, we will analyze the transfer of lots between bottleneck work centers of two wafer fabs in the case of a heavy overload in one of the two wafer fabs and the consequences on individual production planning in the two wafer fabs. To the best of our knowledge, such a setting is not studied in the literature so far (Mönch et al. 2018b; Missbauer and Uzsoy 2022). Since planning and control activities in

semiconductor supply chains are hierarchically distributed in a natural way, we design a MAS to model the borderless fab setting and the resulting decision-making activities. It is expected to some extent that next-generation production planning and control systems for wafer fabs will be based on software agents (cf. Chien et al. 2016). The production planning and control activities are carried out in a rolling horizon setting using a simulation model of a simplified semiconductor supply chain. The cloud-based infrastructure proposed by Herding and Mönch (2022) will be used for the performance assessment activities.

The paper is organized as follows. In the next section, we will describe the problem at hand and discuss related work. The MAS for the borderless fab scenario is described in Section 3. The results of simulation experiments will be presented in Section 4. Conclusions and future research directions are indicated in Section 5.

2 PROBLEM SETTING

2.1 Borderless Fab Setting and Planning Problem

For the sake of simplicity, only two wafer fabs are assumed in this research. The first one is called the delivering wafer fab. It has a heavily overloaded bottleneck work center. We assume that the corresponding work center of the second wafer fab, the consuming wafer fab, is not overloaded. If the number of lots in front of the bottleneck of the delivering wafer fab n exceeds a threshold Δ at a certain point in time then $n - \Delta$ lots are transferred from the bottleneck work center of the delivering wafer fab to an appropriate work center of the consuming fab. The exchanged lots are the most urgent ones, the ones with the smallest local due dates. These lots are then processed on the machines of the target work center of the consuming wafer fab. After processing in the consuming wafer fab, the lots are transferred back to the delivering wafer fab 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 wafer fabs. The setting can be generalized to arbitrary work centers in addition to the bottleneck work center. Several consecutive process steps are possible to be executed in the consuming wafer fab. The borderless fab scenario is summarized in Figure 1.

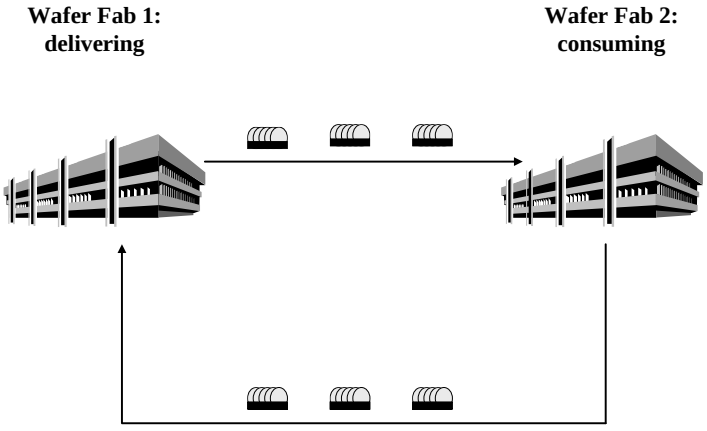

Figure 1. Borderless fab scenario.

We are interested in investigating the impact of production planning in the borderless fab scenario. Production planning is carried out for each of the two wafer fabs using the Simple Rounding Down (SRD) production planning model proposed by Kacar et al. (2013), i.e., we take into account a fixed, exogenous lead time, an estimate of the cycle time of the lots used for planning purposes. We refer to the production planning formulation as SRD since we simply rounding down the known, possible fractional lead time, to an integer multiple of the period length. A cost-based objective function taking into account work in progress

(WIP), inventory, and backlog cost is applied. In the case of the delivering wafer fab, the initial WIP in front of the bottleneck work center must be reduced since lots are transferred to the consuming wafer fab. On the other hand, the initial WIP must be increased in the consuming wafer fab (cf. Section 3). Moreover, it is important that the expected capacity consumption of the transferred lots in future periods is taken into account in the planning model of the delivering wafer fab. Three different production planning settings will be investigated:

1. **Reference scenario with no borderless fab (NBF):** In this reference setting, there is no lot transfer between the two wafer fabs. Production planning will be carried out for each of the two wafer fabs. Production planning is used to adjust the overload situation found in the first wafer fab.
2. **Borderless fab scenario with naive production planning (BF-NPP):** In this setting, lots are exchanged between the delivering and consuming wafer fabs. However, the transferred lots are not taken into account when the production planning models are generated for the two wafer fabs, i.e., the borderless fab scenario is not considered on the production planning level, the lot transfer is only taken into account on the production control level during lot dispatching on the shop floor.
3. **Borderless fab scenario with advanced production planning (BF-APP):** In this setting, lots are again exchanged between the delivering and consuming wafer fabs. This exchange is also considered when the production planning models of the two 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.

We expect that the most advanced BF-APP setting will outperform the first and second one since in this situation the finite capacity is considered in an appropriate way.

2.2 Discussion of Related Work

We discuss related work with respect to borderless fab settings and software agent-based decision support for planning problems in semiconductor supply chains. The idea of exchanging capacity among several close by wafer fabs is considered by Wu and Chen (2007) and Wu and Chen (2008). They propose a simulation-based trading method for two wafer fabs that have established a capacity-sharing partnership for certain work centers. Chien and Kuo (2013) tackle a similar problem by game theory. However, production planning is not considered in these papers. Borderless fab scenarios are discussed by Lendermann et al. (2004) and Gan et al. (2007). They use distributed simulation based on the High Level Architecture (HLA) (cf. SISO, IEEE 2023) that provides a run-time infrastructure (RTI) ensuring the interoperability of the simulation models of the involved wafer fabs. The impact of different lot batching sizes for the cross-fab process step on lot transfer frequency and cycle time is investigated as an application of the distributed simulation approach. Again production planning is not considered.

There are only a few MAS for semiconductor supply chains are described in the literature. Software agents can be seen as software entities that are able to make autonomous decisions to fulfill their own design goals (Weiss 1999). They must have rich communication abilities to support their autonomous behavior. A MAS is a set of agents that interact. The FABMAS system (Mönch et al. 2006b) is designed for scheduling lots in a single wafer fab. An agent-based control approach, i.e. a negotiation approach, for integrated lot and vehicle dispatching is proposed by Wang et al. (2007). The S^{2}CMAS prototype (Herding and Mönch 2016) extends the FABMAS system towards possible applications in semiconductor supply chains. A software agent-based infrastructure for testing and assessing planning approaches for semiconductor supply chains in the cloud is proposed by Herding and Mönch (2022). However, borderless fab scenarios are not supported so far by software agents.

In the present paper, we will design a MAS approach based on the infrastructure from Herding and Mönch (2022) to conduct simulation experiments with production planning formulations for two wafer fabs in a rolling horizon setting. We reuse and extend the S^{2}CMAS MAS prototype in this research. A MAS

approach is applied since the planning problems in semiconductor supply chains are hierarchically distributed (Mönch et al. 2018a). In the borderless fab setting, the management of different wafer fabs might have their own preferences and objectives. This is better supported by a software agent-based decision support where, for instance, also negotiation-based planning approaches are possible (cf. Dudek 2008; Heyne and Mönch 2011; Wu and Chang 2007, 2008).

3 MAS FOR THE BORDERLESS FAB SETTING

3.1 Agentification

The S^{2}CMAS prototype is based on the Product Resource Order Staff Architecture (PROSA) (Van Brussel et al. 1998; Van Belle et al. 2012). PROSA distinguishes decision-making agents (DMAs) from staff agents (SAs). SAs are used to support DMAs by solving their decision problems, and typically encapsulate algorithms for decision making. Distinguishing DMAs from SAs allows separating alternative planning algorithms from the planning system itself. Different solvers, for instance, for mathematical programming or even simulation tools can be incorporated into SAs. In the S^{2}CMAS prototype, SAs use different web services as encapsulation technique for specific planning approaches. To support the borderless fab scenario by the MAS, we first have to create an additional web service which is able to execute the generated SRD production planning linear programming (LP) instance. Next, we have to create additional agents to support the borderless fab scenario. In the borderless fab setup, the agents summarized in Table 1 are important where the bold ones are additional agents compared to the ones of the S^{2}CMAS prototype. The abbreviation DMA refers to decision-making agents while SA indicates that a staff agent is used.

Table 1: Relevant agents of the S^{2}CMAS prototype.

Agent	Type	Description
Borderless fab agent	DMA	- encapsulates the wafer fab overload detection logic - decision making for lots to be transferred - exchanges information with the fab agents
Fab agent	DMA	- coordinates the fab planning agent - coordinates the work area agents - coordinates lot releases -decision making for lot-based decomposition schemes
Fab planning agent	SA	- prepares to execute a specific production planning algorithm - runs the algorithm for a single wafer fab - provides release plans for single wafer fabs
Work area agent	DMA	- coordinates the work of the corresponding work area scheduling agent - decision-making in form of choosing appropriate work area scheduling parameters
Work center agent	DMA	- implements the work area schedules - provides information about a specific work center, e.g. queue length in front of the work center and utilization of the work center

The agents in Table 1 exploit the fact that a single wafer fab can be decomposed into several work areas. Each work area consists of different work centers. A single work center is formed by machines that provide the same functionality, i.e. parallel machines. The fab agent is responsible for decomposing the overall scheduling problem for a single wafer fab into a series of scheduling problems for work areas. Start and completion dates for the different work areas are assigned to each single lot using lot planning

algorithms (cf. Mönch et al. 2013). Based on the additional production planning and control functionality for borderless fab situations described in Section 2, the new borderless fab DMA type is identified for the S²CMAS system. The resulting DMAs are responsible for encapsulating the borderless fab logic as well as selecting and performing the lot transfer between the wafer fabs.

3.2 Modeling of the Lot Transfer Between the Wafer Fabs

The borderless fab agent encapsulates the fab overload detection logic. The work center agent belonging to the bottleneck work center continuously sends the queue length and the utilization of the work center to the borderless fab agent. When an overload occurs, the DMA first decides which lots have to be delivered to the consuming fab (to its fab agent). After that, the process steps which have to be executed by the consuming fab are determined. Finally, the determined lot data is sent to the fab agent of the consuming fab. The fab agent immediately releases the received lots into the base system.

The fab planning agent of the S²CMAS prototype is extended by model generation and modification capabilities. The fab planning agent of the consuming fab receives the message from its fab agent and interprets the transferred data of the message. Next, the SRD planning instance, generated by the fab planning agent, has to be enriched by using the received data. More specifically, the received lots from the delivering fab create new temporary routes, temporary process steps, and temporary products in the base system of the consuming fab which results in changes of the capacity usage. The capacity usage changes are mainly reflected as initial WIP in the current SRD planning instance since the received lots are already released. More data regarding capacity consumption and other required information to generate an SRD problem instance are used from the master data of the consuming fab. The required extensions of the base system are called temporary because if the lots are completed and sent back to the delivering fab, they are no longer necessary. Finally, the generated SRD planning instance will be transferred to the web service where it will be executed. The computed production plan is sent back to the fab planning agent where it is further processed (see Herding and Mönch 2016 for more details). The borderless fab scenario and the related agent interactions are shown as a Unified Modeling Language (UML) sequence diagram in Figure 2.

The borderless fab agent keeps track of all lots which are currently processed in the consuming wafer fab. When a new production plan is required in the delivering wafer fab (in the time span between Steps 4 and 14 of Figure 2), the fab planning agent of the delivering wafer fab requests capacity information of the transferred lots by the borderless fab agent. This capacity information is mainly relevant for determining when the transferred lots will be sent back to the delivering wafer fab since then, more capacity is required in order to continue producing the previously transferred lots on-time (Steps 14-17).

3.3 Required SA for Production Planning

We assume that a planning window of length T with equidistant periods labeled by $t = 1, \ldots, T$ exists. We use the SRD planning formulation proposed by Kacar et al. (2013) as production planning approach in the fab planning agent. The formulation can be stated as follows:

Sets and indices

t:	period index
$g \in G$:	product index for set of all products
$k \in K$:	work center index for set of all work centers
l:	operation index
$O(g)$:	set of all operations of product g
$O(g,k)$:	set of all operations of product g on machines of work center k

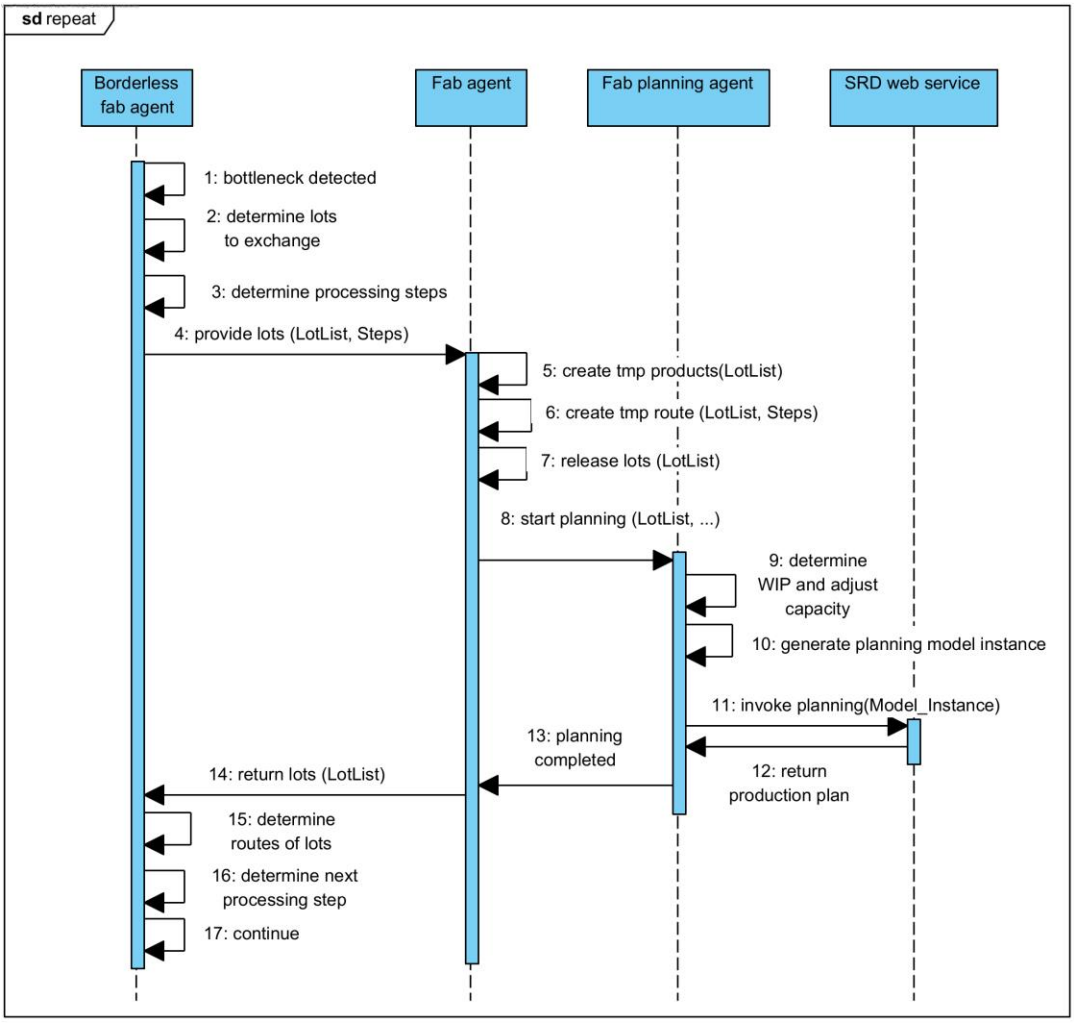

Figure 2: Sequence diagram for borderless fab agent information exchange.

Decision variables

Y_{gtl}: quantity of product g completing operation l in period t

Y_{gt}: output of product g in period t from the last operation of its routing

X_{gt}: quantity of product g released into the first work center of its routing in period t

$\mathrm{W_{gt}}$: WIP of product g at the end of period t

I_{gt}: inventory of product g at the end of period t

B_{gt}: backlog of product g at the end of period t

Parameters

h_{gt}: unit holding cost for product g in period t

ω_{gt}: unit WIP cost of for product g in period t

b_{gt}: unit backlog cost for product g in period t

D_{gt}: demand of product g during period *t*

$Y_{gt}^{(i)}$: initial quantity (in lots) of product g to be completed at the end of period *t*

C_{kt}: available capacity of work center k during period t

α_{gl}: processing time of operation l of product g

L_{gl}: lead time (in number of periods) for product g from release of the raw material to the completion of operation l.

Next, the SRD model can be stated as follows:

$$\min \sum_{g\in G} \sum_{t=1}^{T} \left[\omega_{gt} W_{gt} - h_{gt} I_{gt} - b_{gt} B_{gt}\right] \tag{1}$$

subject to

$$W_{g,t-1} + X_{gt} - Y_{gt} = W_{gt} \qquad t = 1, \dots, T, g \in G \tag{2}$$

$$Y_{g,t} + I_{g,t-1} - I_{gt} - B_{g,t-1} + B_{g,t} = D_{gt} - Y_{g,t}^{(i)} \qquad t = 1, \dots, T, \quad g \in G \tag{3}$$

$$\sum_{g\in G} \sum_{l\in O(g,k)} \alpha_{gl} Y_{gtl} \le C_{kt} \qquad t = 1, \dots, T, k \in K \tag{4}$$

$$Y_{gtl} = X_{g,t-\lfloor L_{gl} \rfloor} \qquad t = 1, \dots, T, g \in G, l \in O(g) \tag{5}$$

$$X_{gt}, Y_{gtl}, Y_{gt}, W_{gt}, I_{gt}, B_{gt} \ge 0 \qquad t = 1, \dots, T, g \in G, l \in O(g). \tag{6}$$

The objective (1) seeks to minimize the cost, i.e. minimize WIP, inventory, and backlog costs. Constraints (2) represent the WIP balance. Constraints (3) are inventory balance equations. The capacity restrictions for each work center are ensured by constraints (4). Integer lead times that are a multiple of the period length are incorporated into the model by the input-output relation constraints (5). The range of the decision variables is modeled by the constraints (6). The lead time quantities are determined by the recursion

$$L_{gl} \coloneqq L_{g,l-1} + FF_g \alpha_{gl}, g \in G, l = 2, \dots, \lfloor O(g) \rfloor, \tag{7}$$

where the values of the flow factors (FF) FF_g for product g are determined by long simulation runs for a prescribed bottleneck utilization. The recursion (7) is initialized by $L_{g0} \coloneqq 0$. Integer lead times are obtained by rounding these values down to an integer multiple of the period length. For a more detailed description of the SRD model, we refer to (Kacar et al. 2013).

When the fab agent receives a message from the borderless fab agent, the agent needs to modify parts of the base system. Figure 3 shows an exemplified lot information exchange request which takes place during the request associated with Step 4 of Figure 2.

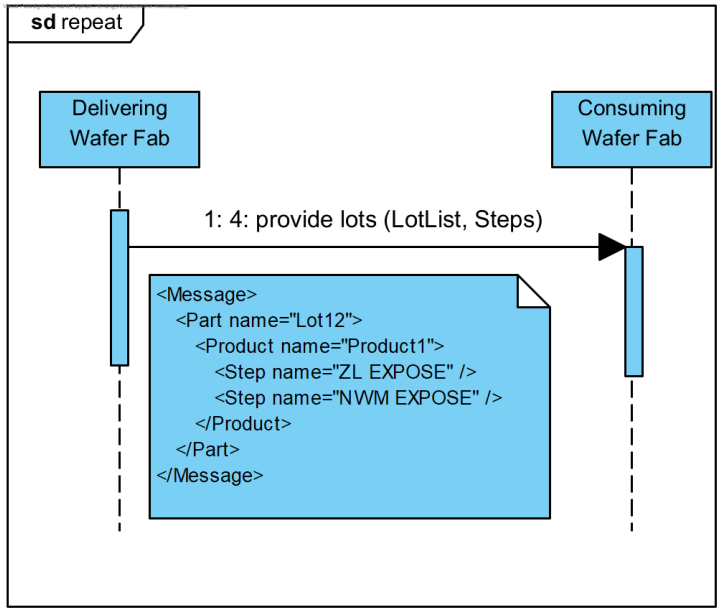

Figure 3. Example for lot information exchange request.

When the fab agent of the consuming wafer fab interprets this message, a temporal product named "tp_Product1" with a route "tr_Product1" is created within the base system of the consuming wafer fab. The processing steps are assigned to the route in the order contained in the message. The step name must match with already existing processing steps of the consuming wafer fab. The lots are immediately released into the base system.

When a new production plan is required in the consuming wafer fab, the fab agent communicates with the fab planning agent. The fab planning agent is able to generate a problem instance of the SRD model. The previously received and released lots are included as initial WIP values W_{g0} and via the $Y_{gt}^{(i)}$ quantities. The C_{kt} values are adjusted in such a way that the initial WIP is taken into account.

3.4 Implementation Issues

3.4.1 Architecture

The S^{2}CMAS prototype is implemented in the C# programming language by extending the FABMAS system (Mönch et al. 2006b) that is based on the ManufAg framework (Mönch et al. 2006a). ManufAG allows for implementing distributed hierarchically organized MAS. The used web services are stateless. They are coded in the C# programming language. The communication between the agents of the MAS and the web services is based on the HTTP protocol.

Discrete-event simulation is used to assess the MAS where the planning functions are applied in a rolling horizon setting. The center point of the proposed architecture is a blackboard-type data layer coded in the C++ programming language in the memory of the simulation computer that is between the S^{2}CMAS prototype and a simulation model of the base system of the semiconductor supply chain. It contains all the relevant business objects such as lots, machines, and products with corresponding routes. These objects are

updated whenever status changes occur in the simulation using the notification mechanism of the commercial simulation engine AutoSched AP.

3.4.2 Deployment of the MAS

The MAS is deployed in the cloud-based environment proposed by Herding and Mönch (2022). The infrastructure allows a scalable deployment of an Autosched AP based simulation which uses the S^{2}CMAS prototype. The described base system represented by the Autosched AP simulation instance as well as the blackboard-type data layer are deployed into a single instance in the cloud environment. A single instance is sufficient due to its fairly low computing and system requirements. The used web services are deployed behind a load balancer which enables auto-scaling.

4 SIMULATION EXPERIMENTS WITH THE MAS

4.1 Simulation Models

A slightly simplified version of the semiconductor simulation testbed proposed by Ewen et al. (2017) that is publicly available as Testbed (2023) is used. It consists of two identical wafer fabs, each with more than 200 machines. Two products with more than 200 process steps are processed on more of 200 machines that are organized in around 80 work centers. The models contain batch processing machines and highly reentrant process flows. Exponentially distributed machine breakdowns are used in the simulation models of the testbed. The cycle time of the two products is between two and three weeks depending on the utilization of the planned bottleneck which is the stepper work center. First-in-first-out (FIFO) dispatching is used at all work centers.

4.2 Design of Experiments

The lot transfer scheme and production planning models are carried out in a rolling horizon setting using discrete-event simulation. A simulation horizon of a single years is applied together with a planning window that consists of 26 periods each of them with a length of a single day.

Normally distributed stationary demand is used that results in 93% - 96% planned bottleneck utilization (BNU) in the delivering wafer fab. 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%. Demand is generated in a similar way as for the delivering wafer fab that is appropriate for this BNU value. All process steps of the transferred lots from the bottleneck work center are performed in the consuming wafer fab. The threshold value $\Delta = 32$ is applied. Local due dates of the lots are set using lot planning with a prescribed FF value.

We are interested in maximizing the profit, i.e. the difference of revenue and the sum of WIP, inventory, and backlog costs. We apply b_{gt}:=50, ω_{gt}: =20, and h_{gt}: = 15 in the simulation experiments. Moreover, the revenue per lot is $r_{gt} := 180$. Five independent demand instances are used in the simulation experiments. Moreover, five 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.

We are interested in assessing the performance of the three settings described in Subsection 2.1., namely the NBF, the BF-NPP, and the BF-APP where we expect that the BF-APP will lead to the largest profit. NBF is used as base line whereas the profit of the BF-NPP and the BF-APP will be reported relative to the one of the NBF scheme. The design of experiments is summarized in Table 2. Note that we solve a total of 27375 LP instances due to the rolling horizon approach.

Table 2: Design of experiments.

Factor	Level	Count
BNU	high	1
CV of demand	0.25	1
Production planning settings	NBF, BF-NPP, BF-APP	3
Independent demand realizations		5
Independent simulation replication per demand realization		5
Total number of simulation runs		75

4.3 Simulation Results

We observe from the computational results that the profit is increased by around 7.4% (compared to the NBF profit) in the case of the BF-NPP setting. This seems reasonable since lots will be transferred from the delivering wafer fab to the consuming wafer fab. The delivering wafer fab has a very high utilization which leads often to several bottleneck situations. On the one hand, these overload situations can often be mitigated by transferring the most urgent lots to the consuming fab. The consuming wafer fab, on the other hand is not so highly utilized which enables the wafer fab to consume the received lots without observing large reductions in profit.

The BF-APP setting, however, leads to a profit increase of up to 11.3% compared to the NBF setting. Compared to the BF-NPP setting, the BF-APP setting incorporates the capacity consumption of the received lots into the production planning instance. An accurate capacity modeling within the production planning function leads to a more accurate release plan which increases the profit. The same is true for the production planning instance of the delivering wafer fab. Incorporating the return time of the lots into the capacity view results in an even more accurate release plan.

5 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this paper, we discussed borderless fab scenarios for two wafer fabs. The design and the implementation of a corresponding MAS were described. The MAS approach allowed to assess production planning formulations for each of the wafer fabs in a rolling horizon setting. We demonstrated by simulation experiments that it is worth to exchange lots between the two wafer fabs. Moreover, we also observed by the simulation results that a detailed modeling of the transfer of the lots between the two wafer fabs is beneficial with respect to capacity when the production planning models are generated.

There are several directions of future research. First of all, we believe that more general borderless fab situations need to be considered, i.e., more work centers in the delivering and consuming wafer fabs and even more than two wafer fabs can be considered. As a second research avenue, it would be desirable to fully automate the generation of the resulting LP models by means of an ontology in such a generalized setting. 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 the static rule-based approach used in the present paper. We also expect that the borderless fab setting must be considered in network-wide approaches such as master planning since the boundaries between single frontend nodes are blurred in such a setting.

It is also interesting to test negotiation approaches, for instance, for sharing capacity among the different wafer fabs using the proposed MAS prototype.

ACKNOWLEDGMENT

The authors were supported by the SC3 project which receives funding from the ECSEL JU under grant agreement No 101007312.

REFERENCES

Chien, C.-F., H. Ehm, J. W. Fowler, and L. Mönch. 2016. "Modeling and Analysis of Semiconductor Supply Chains (Dagstuhl Seminar 16062)". *Dagstuhl Reports* 6(2): 28-64.

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.

Dudek, G. 2008. *Collaborative Planning in Supply Chains: A Negotiation-based Approach*. 2nd ed., Berlin: Springer.

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.

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. 2016. "S^2CMAS: An Agent-based System for Planning and Control in Semiconductor Supply Chains". In *Proceedings MATES 2016*, LNAI 9872, September 27th – 30th, Klagenfurt, Austria, 115-130.

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, 2022.

Heyne, D., and L. Mönch. 2011. "An Agent-based Planning Approach within the Framework of Distributed Hierarchical Enterprise Management". *Journal of Management Control* 22(2): 205-236.

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.

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 *Proceedings of the 2004 Winter Simulation Conference*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. J. Buckley, and J. A. Miller, 1896-1901. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Missbauer, H., and R. Uzsoy 2020. *Production Planning with Capacitated Resources and Congestion*. 1st ed., New York: Springer.

Missbauer, H., and R. Uzsoy. 2022. "Order Release in Production Planning and Control Systems: Challenges and Opportunities". *International Journal of Production Research* 60(1):256-276.

Mönch, L., J. W. Fowler, and S. J. Mason. 2013. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*. 1st ed., 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 M. Stehli. 2006a. "ManufAg: A Multi-agent-system Framework for Production Control of Complex Manufacturing Systems". *Information Systems and e-Business Management* 4:159-185.

Mönch, L., M. Stehli, J. Zimmermann, and I. Habenicht. 2006b. "The FABMAS Multi-Agent-System Prototype for Production Control of Waferfabs: Design, Implementation, and Performance Assessment". *Production Planning & Control* 17(7):701-716.

SISO, IEEE. 2023. High-Level Architecture (HLA), http://www.sisostds.org/. Accessed 2nd May 2023.

Testbed. 2023. http://p2schedgen.fernuni-hagen.de/index.php?id=296. Accessed 2nd May 2023.

Van Belle, J., J. Philips, O. Ali, B. Saint Germain, H. Van Brussel, and P. Valckenaers. 2012. A Service-Oriented Approach for Holonic Manufacturing Control and Beyond. In *Service Orientation in Holonic and Multi-Agent Manufacturing Control*, edited by T. Borangiu, A. Thomas, and D. Trentesaux, 1-20, New York: Springer.

Van Brussel, H., J. Wyns, P. Valckenaers, L. Bongaerts, and P. Peeters. 1998. "Reference Architecture for Holonic Manufacturing Systems: PROSA". *Computers in Industry* 37(3):225-276.

Wang, K.-C., Lin, J. T., and G. Weigert. 2007. "Agent-based Interbay System Control for a Single-loop Semiconductor Manufacturing Fab". *Production Planning and Control* 18(2):74-90.

Weiss, G. (ed.). 1999. *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence.* Cambridge, Massachusetts: MIT Press.

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.

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 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. His email address is lars.moench@fernuni-hagen.de.