Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

AGGREGATED HIERARCHICAL MODELING AND SIMULATION IN SEMICONDUCTOR SUPPLY CHAINS

Georg Laipple
Marcin Mosinski

Robert Bosch GmbH
Tübinger Straße 123
Reutlingen, 72762, GERMANY

Oliver Schönherr
Elias Winter

Saxony Media Solutions GmbH
Bärensteiner Straße 27
Dresden, 01277, GERMANY

Kai Furmans

Karlsruhe Institute of Technology (KIT)
Gotthard-Franz-Straße 8
Karlsruhe, 76131, GERMANY

ABSTRACT

The semiconductor sector is undergoing one of the fastest market growths. Demand is increasing and market forecasts are optimistic. New markets are emerging and product portfolios are broadening significantly. Dynamic supply chains are developing with increasing number of customers, products, suppliers and manufacturing partnerships. Up to now due to modeling complexity and computation time constraints, disjoint systems are used for local supply chain control and optimization. For efficient control, these complex semiconductor supply chains require a global approach for simulation and optimization. This is the motivation for this conceptual paper. In the conceptual part this paper introduces and discusses novel model aggregation approaches and model validity improvements with novel hierarchical modeling concepts. In the modeling part this paper incorporates that into an aggregated simulation model approach as well as a novel hierarchical interface concept for coupling of disaggregated analytical and simulation models to systematically improve overall model validity at Bosch.

1 INTRODUCTION TO SEMICONDUCTOR SUPPLY CHAINS

A semiconductor supply chain is an extremely complex and dynamic system. It involves multiple production partners and several manufacturing stages. The general semiconductor supply chain contains the following consecutive manufacturing stages: Wafer fab, Wafer test, DieBank, Dicing, Assembly, Packaging, Final test (e.g. Mönch et al.2013; Atherton and Atherton 1993). Principally this sequence can be divided to a frontend section and the backend section. Within every production facility (fab) there are up to 1000 different machines and different products. Due to the diversity of products most facilities in frontend are organized in job shop production systems. Assembly and packaging facilities are organized in flow shops. The frontend part of the supply chain contains all processes on wafer (Wafer fab, Wafer test) while backend production units after dicing are chips. Due to the job shop character with up to 1000 different process steps a product can spend a cycle time of several months to complete the frontend part of the supply chain. (Mönch et al. 2013)). The DieBank separating push-oriented Frontend from pull-oriented Backend is used as a decoupling stock. The cycle time in backend is thereby much lower than in frontend. Modeling the whole frontend and backend supply chain in one detailed simulation model with all hundreds of products and thousands of process steps with acceptable calculation time is a hard issue. But because of

978-1-5386-6572-5/18/$31.00 ©2018 IEEE

the complexity and diversity in production systems, synchronization of semiconductor supply chains in a global model is inevitable. The lack of a model capturing the whole semiconductor supply chain in acceptable granularity and calculation time is the motivation of this paper.

For this conceptual paper two objectives are set to close this gap:

1. Develop systematic strategies to aggregate detailed information to be able to model the whole supply chain on an aggregated level. Modelers have to be enabled to describe a facility in the aggregated supply chain equivalent to a machine within a facility.
2. Develop hierarchical modeling concept to couple independent detailed local micro simulation models to the aggregated model. Thereby highly granular models describing critical bottleneck stages of the supply chain can be modularly coupled to the aggregated model to improve the overall model validity.

To specify the gap of an aggregated hierarchical supply chain simulation model in the following section an overview about relevant publications in this area is given.

2 LITERATURE REVIEW OF SUPPLY CHAIN MODELING

Modeling in supply chains has evolved significantly in the recent years. According to Levalle (2018) supply chains integrate a large number of interconnected facilities which dynamically interact and update their behavior policies based on the environment changes. There are various approaches to model these complex supply chains.

Simulation, especially discrete event simulation (DES), is an often used modeling approach in manufacturing systems (e.g. Banks 1998; Cassandras and Lafortune 1999; Law and Kelton 2000; Kohn et al. 2009). Other simulation approaches like system dynamics and agent-based simulation are also fairly often applied in manufacturing systems (e.g. Kádár et al. 2005). Campuzano and Mula (2009) differentiate between various simulation approaches for supply chains like spreadsheet, system dynamics, discrete event simulation (DES) and business games. Zheng et al. (2008) present an overview of simulation approaches in supply chains. Herding and Mönch (2016) develop an agent-based simulation approach for semiconductor supply chains. In this agent-based prototype, unlike DES environments, the agents of the supply chain choose control policies independently and communicate via web services. Kádár et al. (2005) combine the advantages of DES with agent-based simulation within one manufacturing simulation environment. Campuzano and Mula (2011) develop a system dynamics approach for supply chain performance improvement. Stäblein et al. (2007) combine a continuous simulation approach for supply chains with components from artificial intelligence. Rohde (2004) uses an artificial intelligence approach with neural networks to anticipate supply chain outcome. Schodl (2009) combines analytical models within a simulation approach. Heckmann (2016) combines simulation with queueing theory. Chiadamrong and Piyathanavong (2017) are improving supply chain models with a hybrid analytical and simulation modeling approach. Daganzo (2003) describes an aggregated view of the supply chain in an analytical approach.

To reduce calculation time of complex simulation models Roy and Arunachalam (2004), Fujimoto (1993) as well as Campuzano and Mula (2009) concentrate their works on parallel and distributed simulation for supply chains. Chien et al. (2011) reflect that important research in modeling and analysis of semiconductor supply chains consists in the granular description of the entire supply chain for planning and scheduling reasons as well as in the development of distributed simulation techniques to be able to capture entire supply chains of high complexity on a granular level. But all modeling approaches vary between model validity respectively model aggregation and computation time. Computation time constraints motivate modelers either for high model abstraction to cover a broad range of the supply chain or for limitation of modeling range to allow granular models. The lack of a modular hierarchical supply chain simulation model covering broad supply chains with a high granularity is the motivation for this paper.

As most simulation methods already exist, in the following sections the authors are focusing on adequate aggregated description of the supply chain. Suitable aggregated supply chain description will stabilize model validity while reducing calculation effort.
Additionally a hierarchical coupling method will be introduced to improve the aggregated model with the hierarchical coupling of underlying granular models. With this hierarchical method model validity can be improved and computation time will be kept acceptable.

3 AGGREGATION STRATEGIES

As we are not able to describe the behavior of the whole supply chain in a model that captures every production parameter of every physical machine and worker we have to aggregate the behavior.
The aggregation will simplify the model and thereby reduce calculation time but also model validity. From the physical structure semiconductor supply chains can be described in three different aggregation levels:
The **topmost level** confines itself to the intuitive description of collaboration of plants within a supply chain. On this level single plants are described as black boxes with stochastic processing time. In semiconductor supply chains the topmost level is a simple description of the three stages, frontend, backend and customer. In this sequential three-stage-supply chain the customer announces periodic demand which is directly passed to the backend and forwarded to the frontend. The demand can directly or indirectly trigger periodic releases to frontend and backend (Daganzo 2003). The topmost level helps describing very rough, long-term scenarios in huge supply chains beyond the three semiconductor stages. It aims to support strategic, long-term decisions for load balancing and capacity planning in huge global supply chains. This paper only refers to the semiconductor part of the supply chain (frontend, backend and customer). Thereby the topmost level will not be considered because for this purpose it is too high aggregated.

The **macro level** describes the network of facilities within the frontend and backend triggered by customer demand in higher granularity. The use case semiconductor supply chain consists 12 different facilities. These are part of frontend and backend and can be clustered in production, stock and test facilities. Thereby the macro level is much more detailed than the topmost supply chain level. But it still follows a strict aggregation strategy as it does not capture the behavior of the single machines and workers within facility. Models on the macro level can support mid-term decisions in capacity planning and capacity utilization planning, release planning and load balancing as well as work force qualification and workforce assignment.

The lowest level, the **micro level**, enables modeling of local short-term scenarios like short-term delivery forecasts, maintenance scheduling, local OEE improvements. Modeling the complete supply chain on the micro level would result in high model validity on the one hand but also unacceptable calculation time on the other hand. This issue motivated the authors of this paper to derive specific aggregation strategies for supply chain simulation models. The aggregation strategies are basing on the three hierarchical supply chain levels described above and will be introduced below. The objective of this paper is to describe a single facility in the aggregated supply chain equivalent to a single machine in a facility in the existing micro simulation models. The term machine describes a physical equipment, necessary to produce goods. The term facility describes a physical shop floor, factory or fab where multiple machines are located and managed by a shift workforce. Therefore in the following section the aggregation of machine states and bottlenecks will be derived. Subsequently aggregation strategies for cycle time, process time and products will be described.

3.1 Aggregation of Machines

A single semiconductor machine can cost between 100 thousand and 40 million dollars (Mönch et al. 2013). These huge prices are the reason machine utilization is tried to be maximized. Every machine has to be specified concerning machine behavior, capacity, process time, utilization and throughput.

In general in semiconductor industry the following four different machine models can be specified (e.g. Kohn 2015; Atherton and Atherton 1995; Mönch et al. 2013):

- A single job machine, that operates just one process with one lot at the same time
- A batch machine, that operates one process with more than one lot at the same point in time (e.g. diffusion furnace)
- A parallel machine that has more than one process chamber handled with a common handling unit. Each chamber can operate different processes with different lots semi-independently at the same time. (e.g. sputter machines)
- A cluster machine, that can operate more than one process in a sequence within the same machine (e.g. lithography tools)

Altogether the behavior of a generic machine in semiconductor supply chains must be specified in three dimensions: Height (batch size), width (parallel size) and length (cluster size). On the aggregated macro level the single facility node also has to be specified in the same three dimensions. It can operate different processes at the same point in time equivalent to a parallel machine but with multiple handling units (operators or capacitated automation system). A facility can operate more than one lot at the same process at the same point in time (batch). And a facility operates a product specific sequence of processes independently equivalent to a cluster machine.

To describe a facility equivalent to a machine in a job shop production network it must be modeled as a parallel batch cluster machine without common handling unit.

3.2 Aggregation of Machine State

Single machines can have the following six different states: Productive state, standby state, engineering state, scheduled downtime state, unscheduled downtime state, nonscheduled state (SEMI-E10-0304E (2004). This section determines the actual state of a facility considering different single machine states. The aggregated facility node on the macro level can have multiple states at the same time like a parallel machine. This means the aggregated facility is still available if one machine on the micro level is down. The downtime of a single machine influences the facility according to the importance of the machine. A central hypothesis of this paper is that a facility can be described by a sophisticated description of its critical bottleneck machines (e.g. Zhou and Rose 2009).

Downtimes of highly utilized bottleneck machines are more influencing the performance of the facility than downtimes of lower utilized machines. Either behavior of a facility can be described only by its critical bottlenecks ore by a dynamic combination of all its machines with weighted criticality. Both approaches are elaborated in the subsequent sections. The following section describes a dynamic aggregation of machines states according to machine criticality. Therefor in a first step the machine status has to be weighted with the utilization of the machine. The higher the utilization of the machine, the higher the importance of the machine.

The following formula summarizes the machine state aggregation strategy considering machine utilization:

$$aggregated\ state_m = \frac{1}{\mathrm{N}} * \frac{\sum_{n=1}^{N}(t_{mn} * U_n)}{24 * U_I} \quad (1)$$

m = ϵ {1,…6} (six different machine states according SEMI-E10-0304E (2004))
n = machine index ϵ {1,…N}
t_{mn} = time (in hours) machine_n is in state_m
U_n = Utilization of machine_n
U_I = Utilization of facility_I

With formula (1) the utilization-weighted average machine status is normalized by the overall facility utilization. Thereby the sum of state rates is 1.
The aggregation strategy can be extended in precision if the states are weighted additionally with its flow factor. The flow factor informs about the time spent at the machine processing, concerning overall cycle time spent at the machine.

$$FF_n = \frac{CT_n}{PT_n} \quad (2)$$

FF_n = Flow factor of machine$_n$ (ratio of (value adding) processing time and cycle time)
CT_n = Cycle time of machine$_n$
PT_n = Processing time of machine$_n$

The lower bound of the flow factor is determined as 1 when cycle time equals processing time.
If the flow factor exceeds 1, waiting time is increased and the machine due to any variability issues stows material. For the aggregated facility performance the machines stowing material are more critical as they are the real bottlenecks even if the other machines have the same utilization on average. If a facility is utilized equally with 100% the machine with the highest variability is the most critical for facility performance. The states of this machine must be recognized in higher proportional. This is the reason at in formula (3) the following extension of the state aggregation is proposed:

$$aggregated\ state_m = \frac{1}{\mathrm{N}} * \frac{\sum_{n=1}^{N}(t_{mn} * U_n * FF_n)}{24 * U_I * FF_I} \quad (3)$$

FF_I = Overall flow factor of facility$_I$ (overall CT$_I$ divided by overall PT$_I$

Additionally to consider critical one of a kind machines in the facility in higher proportion, the machine state has to be weighted with the ratio of the machines with other machines of the same kind. The fewer alternative machines in a facility the higher the criticality of the single machine. Low performance of the single machine can no more be buffered. The following formula incorporates the third state aggregation extension:

$$aggreg.state_m = \frac{1}{\mathrm{N}} * \frac{\sum_{n=1}^{N}(t_{mn} * U_n * FF_n * xoak_n)}{24 * U_I * FF_I * \frac{1}{\mathrm{N}} * (\sum_{n=1}^{N} xoak_n)} \quad (4)$$

The goodness of fit of these four conceptual aggregation strategies still has to be validated in the use case semiconductor supply chain.

The following table (Table 1) illustrates the aggregation strategy considering utilization, flow factor and proportion of alternative machines of the same kind.

Table 1 Example machine state aggregation.

	x of a kind machines	Flow factor	Util.	Process.	Standby	Scheduled downtime	Unscheduled downtime
Machine1	0,20	3,4	0,7	1	0	0	0
Machine2	**1,00**	2	0,4	0	0	**1**	0
Machine3	0,33	2	0,4	1	0	0	0
Machine4	0,20	3,1	0,7	1	0	0	0
Machine5	0,13	**7**	**1**	0	0	0	**1**
Machine6	0,25	1,4	0,5	0	1	0	0
Machine7	0,13	**15**	0,9	**1**	0	0	0
Machine8	0,20	4,6	0,4	1	0	0	0
Aggregation1: Unweighted status average				**0,625**	**0,125**	**0,125**	**0,125**
Aggregation2: Utilization-weighted status average			**0,625**	**0,620**	**0,100**	**0,080**	**0,200**
Aggregation3: Utilization-weighted, FF-weighted status average		**3,64875**		**0,709**	**0,024**	**0,027**	**0,240**
Aggregation4: Utilization-weighted, FF-weighted, xoak-weighted status average	**0,635270833**			**0,636**	**0,034**	**0,157**	**0,172**

With this strategy the critical machines (Machine2, 5 and 7) are considered over-proportional. When considering only influences of utilization (Aggregation2) in Table1, the rate of unscheduled downtime state increases due to Machine5. Additionally, considering flow factor in Aggregation3, due to critical Machines5 and 7 both, processing rate and unscheduled downtime rate are increasing. Finally integration of proportion of alternative machines of the same kind (Aggregation4) significantly increases scheduled downtime state because the critical one of a kind machine (Machine2) is actually scheduled down.

Principally, the more critical a machine is the more influence it must have in the aggregated model. It is possible to only describe the facility behavior by its bottlenecks. For this purpose the following section proposes a determination strategy for bottlenecks and subsequent sections will further elaborate this bottleneck approach.

3.3 Determination of Bottlenecks

Introduced by the theory of constraints, complex supply chain performance is mainly determined by its bottlenecks (Zhou and Rose 2009). Considering the pareto machine utilization curve, it can be useful to only consider the topmost limiting machines to describe the performance of a facility. A critical issue is the way of identifying dynamic product mix-dependent bottlenecks within a facility. In this approach a combination of different indicator measures is introduced.

First of all, the utilization ranking can indicate bottlenecks. The topmost X percent utilized machines are proposed to be considered as bottlenecks. The residual (1-X) percent of machines of the facilities are proposed to be recognized by a slowdown factor capturing facility influences by unexpected events from underutilized machines. Effective values for X have been experienced to range between 5% and 20%in the use case semiconductor supply chain.

Considering only utilization maxima for bottleneck identification, especially in facilities with balanced machine utilization, may not be sufficient. For this reason, the machine flow factor is integrated in this approach. Machines with comparably high flow factors are jamming material. These must be captured as

bottlenecks even if other machines have equal or higher utilizations. The reason for comparably high flow factors with equal utilization can arise from the high variability of the machine.
The following formula (5) introduces the proportional machine flow factor:

$$rFF_{nI} = \frac{FF_n}{FF_I} \quad |\text{for n} \in \{1\ldots\text{N}\} \; (5)$$

rFF_{nI} = proportion of flow factor of machine_n in facility_I
FF_n = Flow factor of machine_n
FF_I = Overall flow factor of facility_I

The combination of machine utilization and proportional machine flow factor in the use case supply chain results in stable differentiation of bottlenecks in a facility. More than 90 percent of bottlenecks identified in this approach have been verified by the line control.

The aggregation strategies for the facility behavior described above can be applied alternatively. The only difference is that in the bottleneck approach the facility state is just depending from the identified bottleneck machines.

To aggregate the behavior of a facility not only states of machines but also process times must be aggregated. This will be discussed in the following section.

3.4 Aggregation of Cycle Time and Process Time

The time a lot spends within a facility in semiconductor industry is called facility cycle time. The cycle time consist of all processing, waiting and on-hold time spent on the sequence of operations (processes) within a facility. Since all three components of cycle time depend on several stochastic parameters at a micro level, they must be treated stochastically.

To be able to correlate stochastic cycle time and process time with actual performance indicators like facility availability and utilization a dynamic, lot-independent accumulation is going to be proposed in this conceptual aggregation approach. In principle every machine can provide different time stamps. To generally capture the cycle time and process time of all machines the three common time stamps queue time, track-in time and track-out time are introduced. Although track-in and track-out times are quite rough for a micro simulation level it can be guaranteed to have this data available for any machine. With the purpose of modelling a whole supply chain on the aggregated level the advantage of data consistency exceeds the disadvantage of lack of data granularity.

The aggregated process time using track-in and track-out times of every operation can be described by the sum of all process times within the facility. Therefore the median from all operations j of a certain time period has to be calculated and the medians have to be cumulated along the sequence of operations to get the product specific dynamic processing time. Formally this can be described the following:

$$PT_{Ip} = \sum_{j=1}^{J} med(TO_{jp} - TI_{jp}) \; (6)$$

for j = {1;..;J} operations in facility_I
for p = {1;..;P} products in facility_I
PT_{Ip} = cumulated stable process time of product_p in facility_I
TI_{jp} = Track-in of product_p at operation_j
TO_{jp} = Track-out of product_p at operation_j

The same procedure can be applied to calculate the dynamic process specific cycle time of a facility. The formula (6) is slightly different as cycle time incorporates the waiting time in front of the operation.

$CT_{Ip} = \sum_{j=1}^{J} med(TO_{jp} - QT_{jp})$ where CT_{Ip} is the cycle time of $product_p$ in $facility_I$ and queue time QT_{jp} = $TO_{(j-1)p}$. On-hold time can be aggregated accordingly. Therefore the lot events have to be acquired.

Alternatively to median, arithmetic mean can serve for stable process time aggregation. The advantage of the arithmetic mean is the easy rescalability to bigger samples. The disadvantage is the sensitivity to outliers. Also the historical time period for averaging the operation process times is critical. While too long periods hide process time stochasticity in the aggregated macro model, too short periods incorporate the risk of ignoring necessary operations (e.g. of low-runner products). The period (in this case one week) has to be determined ensuring at least every operation ran once. Otherwise the aggregated process time and cycle time will incorporate NULL process times and the process time will be too short.

All together with this robust aggregation approach the stochastic process times and cycle times of a specific product can be aggregated across the single facilities.

4 AGGREGATED SIMULATION MODEL

The project Productive4.0 intends to map the supply chain at the macro level and at the micro level in order to contrast the validity of the two models. Since the micro model across all machines of all facilities would be too time-consuming in its calculation and preparation, one aggregation approach discussed above, is to only consider bottlenecks. Therefore, the described approach from Section 3.3 is chosen to determine the bottlenecks of the facility. The aim is to be able to achieve almost the same results when considering the bottlenecks as by looking at all machines.

The most complex facility of the use case is a front-end wafer factory FAB1 with almost 3,000 machines. All 12 factories in the application scenario together contain about 9,000 machines. FAB1 is currently being mapped by a classic micro model simulation across all machines. If the amount of bottleneck machines is one third of the 9,000 machines, there would be as many machines for the bottleneck-focused supply chain model as for the micro model of FAB1. Of course, the number of products and the resulting entities in the system would be larger, but it is more likely to increase the main memory load rather than the computation time. In addition, a concept will be presented in Section 5 that will substantially reduce the processing time of the supply chain model.

The process time on machines is determined from the difference between track-in and track-out time as described in Section 3.4 and mapped to the type of machine like single job machine or batch machine as classified in 3.1. Influences such as failure or maintenance that lead to downtime are taken into account according to the description in Section 3.2. In the bottleneck model, bottleneck influences are precisely taken into account. In addition, all processes that a product passes through between two bottlenecks are combined into a "virtual process" due to calculation time constraints. The process time on a virtual non-bottleneck machine is determined from the difference between the queue-time and the track-out time instead of the difference between the track-in and track-out time. The virtual process does not need resources, it starts with the queue time of its first process and ends with the track-out of its last process (e.g. **Fehler! Verweisquelle konnte nicht gefunden werden.**). Stochastic influences such as maintenance, setup, failures or rework are implicitly taken into account by considering the cycle time in the virtual processes. This simplification significantly reduces the complexity of the model. Since the performance-determining pacemaker processes (bottleneck process steps) are still considered in detail, only a small loss of validity is expected. This is possible because stochastic influences after adopting this approach have no impact on the model if they do not affect bottleneck machine capacities.

5 HIERARCHICAL SIMULATION APPROACH

On the basis of the three hierarchical supply chain levels introduced above, critical facilities of the supply chain can be described in an independent lower level micro model. In this section, a hierarchical model coupling concept is introduced which will be implemented in the next months. This concept strategically improves the aggregated simulation with the detailed results from the lower level micro model with nearly constant calculation time. As a result, individual factories can be simulated in great detail or already existing simulation models can be integrated. To hold the computation time nearly constant it should be possible to simulate the different facilities in parallel with the synchronization across the macro model of the aggregated level. This seems possible because the facilities are not as closely networked as the machines of a conventional micro model. While within a factory the machines exchange wafers within real time, this exchange takes place between facilities at discrete changeover times several times a day. In the considered use case the changeover occurs twice a day. Thereby it is possible to simulate the periods between the changeover times in parallel since in these periods no cross-relationships need to be taken into account by the simulation. When a changeover occurs, each of the facilities is given the opportunity to forward material to other facilities.

In the application scenario, 12 facilities should be considered in the supply chain. Each of the facilities can be simulated on its own arithmetic core, and an additional arithmetic core synchronizes the various cores when the changeover times occur. Since the concept requires 13 cores and there are computers with 12 or 24 cores, the least complex two facilities should be simulated on a single core in order to save resources. Because in the use case the most complex facility exceeds the combined two least complex facilities, combining the least complex two facilities on a single arithmetic core results in no penalty in computation time. Assuming that about one third of the machines are bottleneck machines, the simulation of the 12 facilities of the use case could be significantly faster than the conventional micro model simulation of FAB1 of the scenario described in Section 4. While conventional micro model simulators outsource different replication on different cores and the supply chain simulation distributes the facilities to the cores, the different replications of the supply chain simulation are to be distributed to different computers. This procedure is implemented because the number of communication facilities in the supply chain simulation is much higher than the one-time synchronization of the replication and the communication between cores is faster than that between computers.

5.1 Simulation-Simulation Model Coupling

With the concept described in Section 5 it should be possible to hierarchically link different simulation tools calculating different facilities of the supply chain via the macro model. This would have the advantage of being able to use currently existing micro simulation models within the supply chain. Furthermore, special simulators can be used to illustrate individual facilities particularly well. While certain simulators are particularly well suited to imaging the front end, others are better suited to map the flow shop-oriented backend processes of semiconductor supply chain. As just described, the simulations of the individual facilities are synchronized using changeover times defined by the supply chain simulation model. In order to be able to integrate different simulation tools into the hierarchical simulation, a defined exchange format and an interface should be made available. These should be able to exchange themselves with the macro simulation over a multiplicity of simulators to the discrete changeover times. Since only materials are exchanged between the facilities at the changeover times, the interface also only has to capture the exchange of material at changeover times between the facilities. Currently, we develop a proposal for the interface within the Productive4.0 project consortium to be defined to a standard.

5.2 Simulation Analytical Model Coupling

In addition to different simulators, various analytical optimization models like queuing models (e.g. Heckmann 2006), linear programming models or constraint programming models have to be integrated into the supply chain simulation framework for example for the optimization of local lot sizing problems. The calculated scenario will be used during the subsequent simulation of the facility. The concept is similar to the concept described in 5.1 for describing a supply chain node with an independent micro simulation model. The only difference is that in this concept not a node is replaced by an external simulation model, but that an additional model node is inserted, which is represented by the analytical model. The exchange between the nodes takes place again at the change times. If an analytical node is inserted into the model the products will first move to the analytical node and then to the supply chain simulation nodes at discrete changeover times. During the calculation in the analytical model no discrete simulation time passes. In the concrete example, the products migrate to each discrete changeover time from F2 and F1 to AN1 and are given to F3 at the same discrete point in time. AN1 merely enriches the products with information to be controlled in the simulation in F3. For exchange between the analytical model nodes and the model nodes from the supply chain simulation model, the same interface as in the simulation coupling is used.

6 CONCLUSION AND PERSPECTIVES

Responding to the lack of a valid hierarchical long term semiconductor supply chain simulation model, precise aggregation strategies and a hierarchical coupling approach are discussed in this conceptual paper. The aim is to support strategic decision making in semiconductor supply chains focusing for example facility load balancing use cases. Therefore in the conceptual part of this paper a bottleneck determination strategy is introduced. The bottlenecks in this modeling approach are the pacemakers of the supply chain and have to be precisely recognized in the aggregated model. The residual machines can be summarized to one virtual machine with one virtual process time. To increase the validity of the model detailed local simulation and analytical models of critical stages of the supply chain may have to be integrated. Therefore a hierarchical coupling approach is introduced in the modeling part of this paper. To reduce computation time synchronized, distributed computing strategy is discussed on the basis of the considered use case.

In future work when this approach is completely implemented the hierarchical semiconductor supply chain model will be validated against the physical system. In future research this conceptual modeling approach has to be extended to further increase supply chain model validity under constant computation time. Therefore aggregation strategies for workforce, dispatching policies as well as machine capabilities have to be developed. Additionally model validity differences have to be compared between the different hierarchical levels. To use the model for optimization on this basis a simulation-based optimization approach has to be developed.

ACKNOWLEDGMENTS

This work has received funding from the Electronic Component Systems for European Leadership Joint Undertaking under grant agreement No 737459 (project Productive4.0). This Joint Undertaking receives support from the European Union Horizon 2020 research and innovation program and Germany, Austria, France, Czech Republic, Netherlands, Belgium, Spain, Greece, Sweden, Italy, Ireland, Poland, Hungary, Portugal, Denmark, Finland, Luxembourg, Norway, Turkey.

REFERENCES

Almeder, C., M. Preusser, and R. F. Hartl. 2009. “Simulation and Optimization of Supply Chains: Alternative or Complementary Approaches?”. *OR Spectrum* 31(1):95-119.

Atherton, L. F. and R. W. Atherton. 1995. *Wafer Fabrication: Factory Performance And Analysis.* Boston: Kluwer Academic Publishers.

Banks, J. 1998. *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice.* New York: John Wiley & Sons, Inc.

Campuzano, F. and J. Mula. 2011. *Supply Chain Simulation.* London: Springer.

Cassandras, C. G., and S. Lafortune. 1999. *Introduction to Discrete Event Systems.* Norwell, MA: Kluwer Academic Publishers.

Chiadamrong, N. and V. Piyathanavong. 2017. “Optimal Design of Supply Chain Network under Uncertainty Environment Using Hybrid Analytical and Simulation Modeling Approach”. *Journal of Industrial Engineering International December* 13(4):465–478.

Chien, C.-F., S. Dauzère-Pérès, H. Ehm, J. W. Fowler, Z. Jiang, S. Krishnaswamy, L. Mönch, and R. Uzsoy. 2011. “Modeling and Analysis of Semiconductor Manufacturing in a Shrinking World: Challenges and Successes”. *European Journal of Industrial Engineering* 5(3):254–271.

Daganzo, C. F. 2003. *A Theory of Supply Chains.* Heidelberg: Springer.

Fujimoto, R. M. 1993. "Parallel Discrete Event Simulation: Will the Field Survive?". ORSA Journal on Computing 5(1): 213–230.

Heckmann, I. 2016. "Simulation for Supply Chain Analysis". In *Towards Supply Chain Risk Analytics,* edited by I. Heckmann, 165-179. Wiesbaden: Springer Fachmedien.

Herding, R. and L. Mönch 2016. "S2CMAS: An Agent-Based System for Planning and Control in Semiconductor Supply Chains". In , eds. M. Schillo, M. Klusch, J. Müller, H. Tianfield, 115-130. Berlin: Springer Fachmedien.

Kádár, B., A. Pfeiffer, and L. Monostori. 2005. "Building Agent-Based Systems in a Discrete-Event Simulation Environment". In *Multi-Agent Systems and Applications IV,* edited by M. Pěchouček, P. Petta, and L. Z. Varga, 595-599. Berlin Heidelberg: Springer.

Kohn, R., D. Noack, M. Mosinski, Z. Zhou, and O. Rose 2009. "Evaluation of Modeling, Simulation and Optimization Approaches for Work Flow Management in Semiconductor Manufacturing". In: *Proceedings of the 2009 Winter Simulation Conference*, edited by. M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 1592–1600. Piscataway, New Jersey: IEEE.

Kohn, R., 2015. *A Framework for Batch Scheduling with Variable Neighborhood Search in Wafer Fabrication.* München: Univ. der Bundeswehr, Diss., 2015.

Law, A. and W. D. Kelton. 2000. *Simulation Modeling and Analysis*. New York: McGraw-Hill Science.

Lefeber, E. and D. Armbruster 2010. "Aggregate Modeling of Manufacturing Systems". In: *Simulation Modeling and Analysis*, edited by. A. Law and W.D. Kelton, 509-536, New York: McGraw-Hill Science.

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

SEMI-E10-0304E. 2004. *Specification for Definition and Measurement of Equipment Reliability, Availability, and Maintainability (RAM).* Milpitas, CA: SEMI.

Levalle, R. R. 2018. "Supply Networks". In *Resilience by Teaming in Supply Chains and Networks Automation*, edited by R. R. Levalle, 5-17. Cham: Springer.

Rohde, J. 2004. "Hierarchical Supply Chain Planning Using Artificial Neural Networks to Anticipate Base-Level Outcomes". *OR Spectrum October* 26(4):471–492.

Rossi, F., P. van Beek, and T. Walsh. 2006. "Introduction". In *Handbook of Constraint Programming (Foundations of Artificial Intelligence),* edited by J. Hendler, H. Kitano and B. Nebel, 1-3. North-Holland: Elsevier.

Roy, R. and R. Arunacalam. 2004. "Parallel Discrete Event Simulation Algorithm for Manufacturing Supply Chains". *Journal of the Operational Research Society June* 55(6):622–629.

Schodl, R. 2009. "The Best of Both Worlds - Integrated Application of Analytic Methods and Simulation in Supply Chain Management". In *Rapid Modelling for Increasing Competitiveness*, edited by G. Reiner, 155-162. New York: Springer Publishing.

Stäblein, T., H. Baumgärtel, and J. Wilke. 2007. "The Supply Net Simulator SNS: An Artificial Intelligence Approach for Highly Efficient Supply Network Simulation". In *Management logistischer Netzwerke,* edited by H. O. Günther, D. Mattfeld and L. Suhl, 85-110. Heidelberg: Physica-Verlag.

Zheng, X., H. Yu, and A. Atkins. 2008. "An Overview of Simulation in Supply Chains". In *Advanced Design and Manufacture to Gain a Competitive Edge,* edited by X-T Yan, C Jiang, and B. Eynard, 407-416. London: Springer-Verlag.

Zhou, Z. and O. Rose. 2009. "A Bottleneck Detection and Dynamic Dispatching Strategy for Semiconductor Wafer Fabrication Facilities". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 1646–1656. Piscataway, New Jersey: IEEE.

AUTHOR BIOGRAPHIES

GEORG LAIPPLE is Ph.D researcher in supply chain management at Karlsruhe Institute of Technology (KIT). His supervising professor is Prof. Dr. Kai Furmans. Georg Laipple is associated to the Ph.D program of Robert Bosch GmbH. He is project manager of the public co-funded project Productive4.0 and member of semiconductor manufacturing engineering department in the Robert Bosch GmbH in Reutlingen. Before joining the Ph.D program he was logistics engineer of line control and real time dispatching. His e-mail address is georg.laipple@de.bosch.com.

OLIVER SCHÖNHERR is one of the leaders of Saxony Media Solutions GmbH (SaxMS). He holds a Ph.D in Manufacturing Simulation. He is SaxMS project manager for various optimization, planning and simulation projects for Deutsche Bahn AG, Robert Bosch GmbH and Airbus Group. He has years of experience in developing planning and simulation software in the domain of production and logistics. His email address is oliver.schoenherr@saxms.de.

ELIAS WINTER is one of the leaders and the technical head of Saxony Media Solutions GmbH (SaxMS) in Dresden. Together with his team, he developed a framework for simulation based optimization, the core of the SaxMS.APS that is used in numerous projects of various domains in productions and logistic. His interests in current research originate from the complexity and scope of his projects like aircraft maintenance or supply chain optimization. His email address is elias.winter@saxms.de.

MARCIN MOSINSKI is Senior Manager for FAB digitalization at Bosch Semiconductor Cluster in Reutlingen and Dresden (Germany). He received his M.S. degree in computer science from Dresden University of Technology and his PhD at Munich University of the Federal Armed Forces. His work experience includes Big Data analysis for creation of online simulation models for semiconductor industry (Infineon Technologies, BOSCH Group) and design of new lot dispatch strategy by Globalfoundries. His research interests include simulative and analytical forecasting of complex problems in manufacturing facilities and the statistical data analysis. His email address is marcin.mosinski@de.bosch.com..

KAI FURMANS holds the endowed chair of logistics at the University of Karlsruhe (KIT). From 1996 to 2003 he was director of logistics at Robert Bosch GmbH in the division of thermotechnology. From 1994 – 1996 he was head of the research group "Material Handling Systems" at the University of Karlsruhe (KIT). From 1992 to 1994 he was Post-Doc at IBM in Manufacturing Research. His email address is kai.furmans@kit.edu.