*Proceedings of the 2024 Winter Simulation Conference*
*H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.*

# DIGITAL TWIN BASED UNCERTAINTY INFORMED TIME CONSTRAINT CONTROL IN SEMICONDUCTOR MANUFACTURING

Marvin Carl May[1], Lars Kiefer[1], Gisela Lanza[1]

[1]wbk Institute of Production Science, Karlsruhe Institute of Technology (KIT), Karlsruhe, GERMANY

## ABSTRACT

Semiconductor manufacturing, commonly described as a complex job shop, contains product-inherent constraints that amplify dispatching complexities. Most notably time constraints restrict the maximum waiting time between processes inducing the need to control the release of time constraint lots as violations lead to scrap. The proposed approach uses a fab wide digital twin in form of discrete event simulation based on knowledge graph structure and derives uncertainty informed time constraint violation probabilities through rollouts in near real time. A real-world front end semiconductor manufacturing fab serves as an ex-post validation and shows the benefits of the approach.

## 1 INTRODUCTION

The importance and growth of the electronics industry has been for the past few years stimulated by massive, global trends, such as , Artificial Intelligence (AI), Internet of Things (IoT), Natural Language Processing and smart products and the software-defined characteristics of modern systems (May et al. 2022). As a consequence, the electronics industry has become one of the largest industries world-wide as semiconductors are essential components in all industries nowadays (Maleck and Eckert 2017). Embedding circuits into new product generations is critical for dealing with climate change and resource scarcity (Uçar et al. 2020). Therefore, semiconductors enable the achievement of energy efficiency, individual mobility, security, the establishment of Industrial Internet of Things (IIoT) and the application of AI. This has led to fast innovation cycles in semiconductor manufacturing and thus its technology-intensive and capital-intensive nature (Mönch et al. 2011). The technological intensity is manifested in the presence of forming and cutting processes, electro-physical and chemical processes, abrasive processes, surface engineering and metrology augmented by complex machinery and production system organization as well as highly specialized semiconductor design (Mönch et al. 2013). The maximization of the material potential and avoiding the production of faulty products is crucial and, for that reason, semiconductor manufacturing equipment accounts for a major cost driver (Hong et al. 2023). Hence, in so called fabs, short for semiconductor fabrication plants, operate any day for the whole day. To minimize coordination effort and setup times wafers containing Integrated Circuit (IC) chips are packetized into lots of 25 or 50 (Ziarnetzky et al. 2017). Furthermore, it is necessary to often visit several machines to transform semiconductor material into an IC in a layer by layer manner leading to recurrent material flow (Altenmüller et al. 2020). Beyond individual technological complexity the major challenge for semiconductor manufacturing lies in the coordination of this complex job shop (Mönch et al. 2011). Each wafer, a thin slice of semiconductor material, requires up to 800 processing steps each between a few minutes and several hours (Ziarnetzky et al. 2017). Maintaining the yield high is decisive. In addition to processing or design errors a major yield loss consists of contaminated wafers through native oxidation, crystal formation, ion migration or dust deposition (Lima et al. 2021). These impurities pollute the surface and inhibit the designed electrical flow. Thus, semiconductor manufacturing equipment in fabs is operated in clean rooms to reduce contamination (Klemmt and Mönch 2012). Nevertheless, between many process steps wafers can only remain unprocessed for several hours otherwise yield is reduced as the wafers can often not be recovered and have to be scrapped

(Altenmüller et al. 2020). Hence, adhering to these time constraints is of utmost importance (Arima et al. 2015). System behavior is time transient as the overall time wafers spend within a fab can be up to several weeks and months due to the recurrent material flow (Mönch et al. 2013). Controlling the production under time constraints, re-entrant and non-linear material flow given varying process times is challenging (Wang et al. 2018). Additionally, the process is time-consuming and stressful for operators (Lima et al. 2017b). In this volatile environment production control is amplified by human operators that deal with time constraints (Lima et al. 2019). This requires extra, manually made, effort, that is error-prone, inconsistent and does not optimize opportunities. Making use of the real-time fab data can enable overcoming traditional, rule-based approaches that are too rigid (Altenmüller et al. 2020). Establishing an intelligent, digital twin based production control in this complex job, that minimize time constraints violations, can alleviate the current shortcomings. Additionally, reducing these wasteful activities not only contributes to monetary business objectives, operator well-being, but also to sustainability and the achievement of net zero climate goals as wafer fabrication is energy intensive (May et al. 2021).

## 2 LITERATURE REVIEW

Time constraints limit the maximum time for an individual lot between completing processing of operation *A* and starting operation *B*, where *A,B,...* defines the ordered list of operations to be performed on the particular lot (Klemmt and Mönch 2012). As visualized in Figure 1 simple time constraints regard two consecutive operations *A, A+1*, while timelink area constraints regard non-consecutive operations. Through consecutive or nested time constraints the complexity for production control can be increased (Wang et al. 2018). Thus, most studies restrict themselves to simple time constraints (Altenmüller et al. 2020).

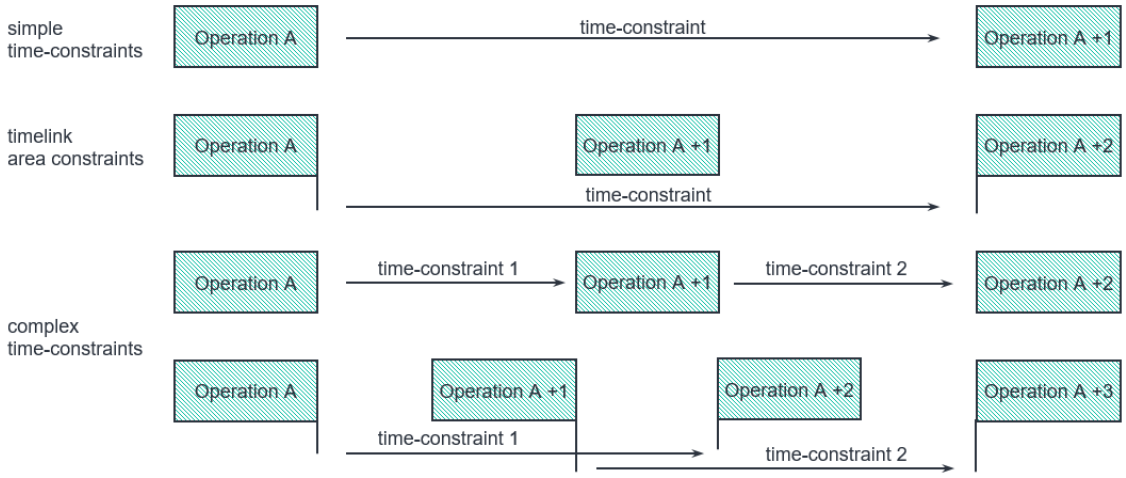


Figure 1: Time constraint types based on (Klemmt and Mönch 2012); (Wang et al. 2018).

Dealing with time constraints by aiming at a minimization of time constraint violations through various levers is a decade old journey (Klemmt and Mönch 2012). It is aggravated with increasingly smaller dimensions that put even more stringent time limits on time constraints. On a capacity planning level (Robinson and Giglio 1999) proposed a queuing theory based time constraint violation probability prediction. Given a specific yield target the optimal number of machines and the associated low violation probability can be obtained. These approaches can be extended to include machine breakdowns (Tu and Chen 2009) or batch equipment (Tu and Chen 2011) and yield a concrete increase in capacity due to the time constraints. As opposed to a simulation approach that aimed at experimentally verifying the influence of different time constraints in semiconductor manufacturing capacity planning, which found that both frequency and time limit have a strong influence on the effects of fab performance (Pappert et al. 2016). A real world machine learning based approach that predicts inventory levels confirmed these results (Chien et al. 2020).

These influences however cannot be appropriately addressed in capacity planning, so that time constraint adherence must be ensured on an operational level (Ono et al. 2006).

Operational control of time constraints is achieved through scheduling or dispatching. The former is widely regarded in literature as it has an abstract, prescriptive nature, is falling in line with traditional fab planning and control as well as the low complexity of dispatching rules applied in real world use cases (Bixby et al. 2006). The main disadvantage of using scheduling, in particular MILP exact solutions, is their limit to two (An et al. 2016), three (Kim and Lee 2017) or in general few machines with 20 to 30 jobs to be scheduled (Yu et al. 2013). Constraint programming can increase the production scale in real world applications (Maleck and Eckert 2017), yet is still limited to areas within a complex job shop (Maleck et al. 2018). Through decomposition these approaches can effectively reduce the average number of time constraint violations (Maleck et al. 2019). To speed up the optimization (Zhou and Wu 2017) propose simulated annealing, that can be extended to regard larger problem settings (Nattaf et al. 2019). However, further extensions with cuckoo search still cannot control entire fabs (Zhou et al. 2019). Oftentimes in semiconductor fab scheduling studies time constraints are not directly regarded but only secondarily profit from reduced cycle times (Lee 2020). One approach in control of a fab uses a related approach that is based on a blacklist and whitelist to select possible lots with time constraints (Winkler et al. 2016). The authors propose a nested simulation model where over a prescribed period individual decisions are taken, which itself are validated in one simulation. The results however are not uncertainty informed as only an individual rollout of a simplified simulation, to avoid the real system complexities, is performed.

On a dispatching level priority rules, heuristics, are the predominant form of controlling material flow in semiconductor manufacturing (Altenmüller et al. 2020). Hence, it is no wonder earliest applications in dispatching study the influence of heuristics and their input to reduce time constraint violations (Scholl and Domaschke 2000). Using these simulations to identify and deduce good heuristics (Zhang et al. 2016), carefully prioritize time constraint lots (Kobayashi et al. 2013) or optimize a rule-based dispatching including criticalities (Kopp et al. 2020) and product-mix (Toyoshima et al. 2013) is widely applied. These studies, however, show, that higher fidelities of simulations give rise to more situational awareness and improved dispatching decisions with respect to time constraint violations (Ciccullo et al. 2014). Within such simulation (Altenmüller et al. 2020) train a deep reinforcement learner to dispatch lots and adhere to time constraint, which is extended by (Valet et al. 2022) who control dispatching and maintenance simultaneously. Thus, the notion of a gate keeping decision in dispatching to avoid dispatching lots with critical time constraints is predominant (Pirovano et al. 2020). For controlling this particular decision heuristic control policies can be derived (Wu et al. 2016). Alternatively, (Sadeghi et al. 2015) present an approach to estimate the time constraint adherence probability by schedule randomization and comparison and evaluation. An improved intelligent sampling approach is presented by (Lima et al. 2017a) and extended to cover full semiconductor fabs (Lima et al. 2017b) and ultimately recognize more complex grouping (Lima et al. 2021). In a similar vein, yet training a machine learning predictor on past real-world fab data, (May et al. 2021) propose uni- and multi-variate time series predictors to evaluate the time constraint violation probability or use future machine tool queue predictions as approximations (May et al. 2021). The extended approach shows significant improvements in avoiding time constraint violations in real-world semiconductor fabs (May et al. 2021). Using simulations not only to validate the proposed dispatching or gate keeping policies but to evaluate the probability of time constraint violations has not been regarded.

Thus, the application of digital twins, up-to-the-minute instantiated high fidelity discrete event simulations of a semiconductor fab, to control time constraint adherence in a semiconductor fab can bridge the gap between simulation and data-based approaches. Thereby the time constraint violation probability refers to an uncertainty aware evaluation of time constraints adherence of an individual lot in the binary domain. This leads to the following two research questions: First, how can a digital twin architecture be derived that can be used to obtain time constraint violation predictions from observation of the digital twin behavior? Second, how can such a digital twin and violation prediction be used to reduce time constraint violations in an actual wafer fab?

## 3 DIGITAL TWIN ARCHITECTURE

A digital twin on a production system level can be described as a discrete event simulation (DES) that can transform manual, intuition based processes to a data-based, clearly described fictional system (Uhlemann et al. 2017). Extending this notion (Negri et al. 2017) envision synchronized simulations on different levels to constitute a digital twin. By evaluating the performance of various control policies in near real-time in the digital twin a foresighted digital twin is constructed (May et al. 2021). Beyond traditional manufacturing and semiconductor fabs the concept of coupling a digital twin with real-time predictions and control has been successfully shown in a petrochemical factory (Min et al. 2019). In the realm of semiconductor fabs the flexibility of the simulation model to accommodate for the large variety of equipment and constraints has shown beneficial (Valet et al. 2022). One particular approach can serve as the blueprint for such a flexible digital twin architecture that is capable of describing such systems, an ontology, knowledge-graph based DES (May et al. 2022). The system is modeled as an instantiated knowledge graph that conforms to an ontology describing the potential interrelations of such a real-world system. Due to this graph like structure the simulation is serialized per se and changes are immediately reflected in the knowledge graph, the core of the simulation. To apply randomization and apply it in large scale in a full semiconductor fab at speed the following digital twin architecture is derived.

### 3.1 DES System Elements

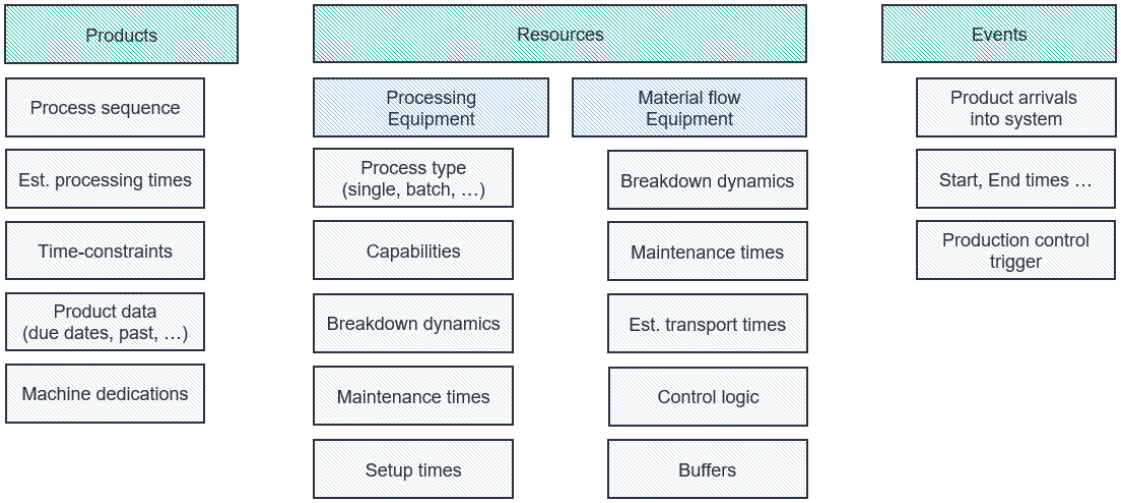


Figure 2: Relevant simulation system elements.

First, based on the OntologySim (May et al. 2022) and previous semiconductor manufacturing system simulations (Kuhnle et al. 2022) required system elements are derived and divided into products, i.e. lots in fabs, resources, in particular processing equipment and transporting equipment relevant for material flow, as well as events to constitute a discrete event simulation. Given expert knowledge and data analysis with respect to queues and time constraints in semiconductor fabs (May et al. 2021), (Mönch et al. 2011) ,(Lima et al. 2021) the simulation system elements are extended as visualized in Figure 2.

The structure of the knowledge graph is based on the ontology used in the OntologySim and extended towards semiconductor specifics. It can be visualized on a general structure of relevant elements for a generic system simulation as in Figure 3. All system elements are hence connected in form of a knowledge graph where vertices describe above mentioned entities and the interrelations are modeled through edges in a similar vein to the OntologySim (May et al. 2022). While process sequence, time constraints and product data can be directly obtained from the fab's database, processing times need to be estimated on

past long-term data. Equipment breakdown dynamics, maintenance and setup times have to be similarly calculated from past data. Regarding the system control either an integration of the existing control system to the digital twin is necessary or the control logic has to be reconstructed based on past behavior (May et al. 2024). Running the simulation model in the DES as a digital twin besides the real-world long-term data requires a short-term, real-time transfer of events to capture the current system status. Due to the knowledge graph based structure a direct transfer from the fab's real-time data can provide the initial starting point for the digital twin.

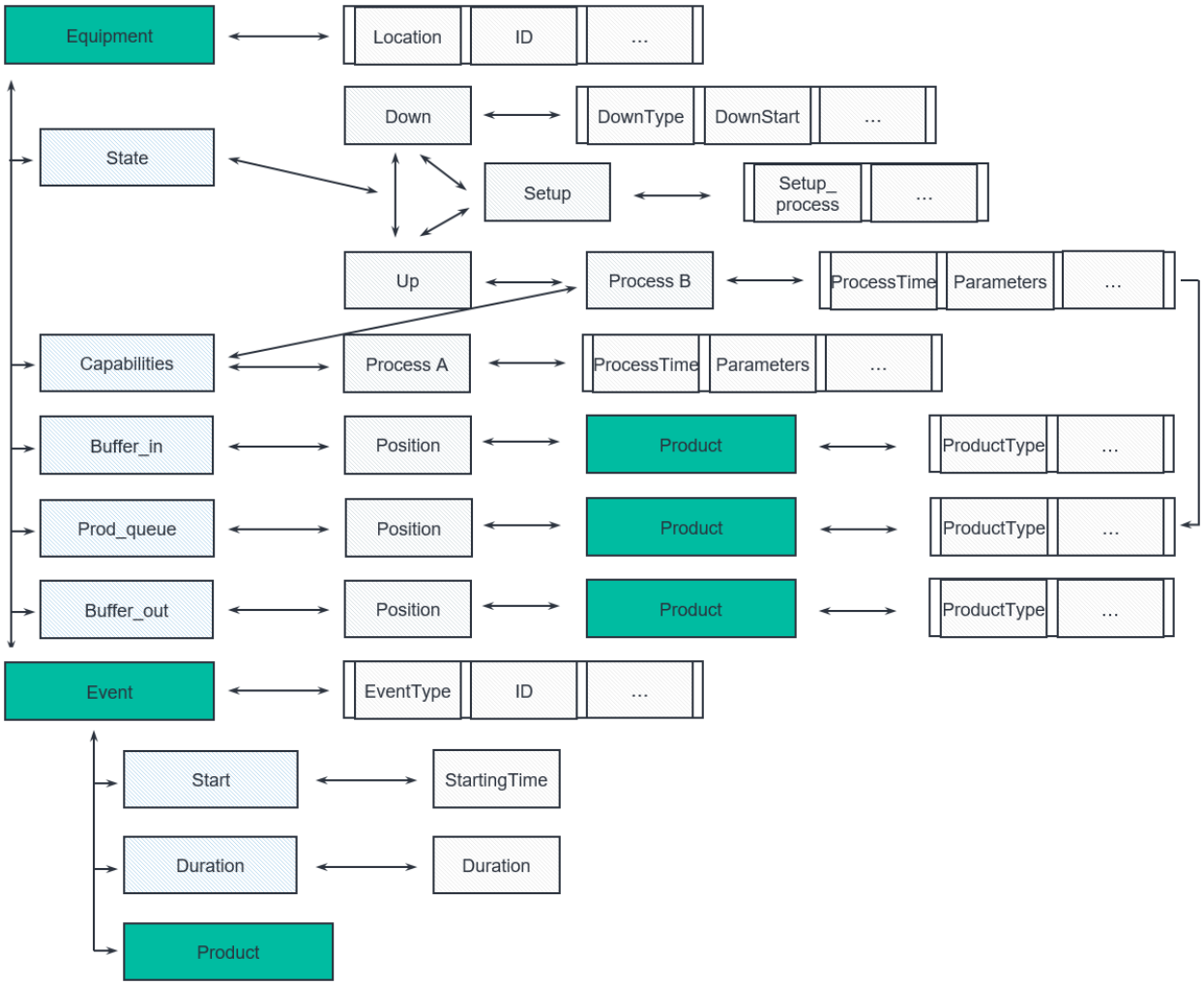


Figure 3: Selection of Knowledge Graph elements and structure based on (May et al. 2022).

### 3.2 Digital Twin Instantiation

To facilitate the digital twin architecture it is desirable to construct a model of the real system that in real time contains the current location, status and control of entities in the knowledge graph based structure that is also used in the DES. This constitutes a real-time data representation, often referred to as the digital shadow, and is used to instantiate the digital twin that can the be used to improve control (May et al. 2021). The overall structure is visualized in Figure 4 and contains the real-time system model as the digital representation which is instantiated to the digital twin.

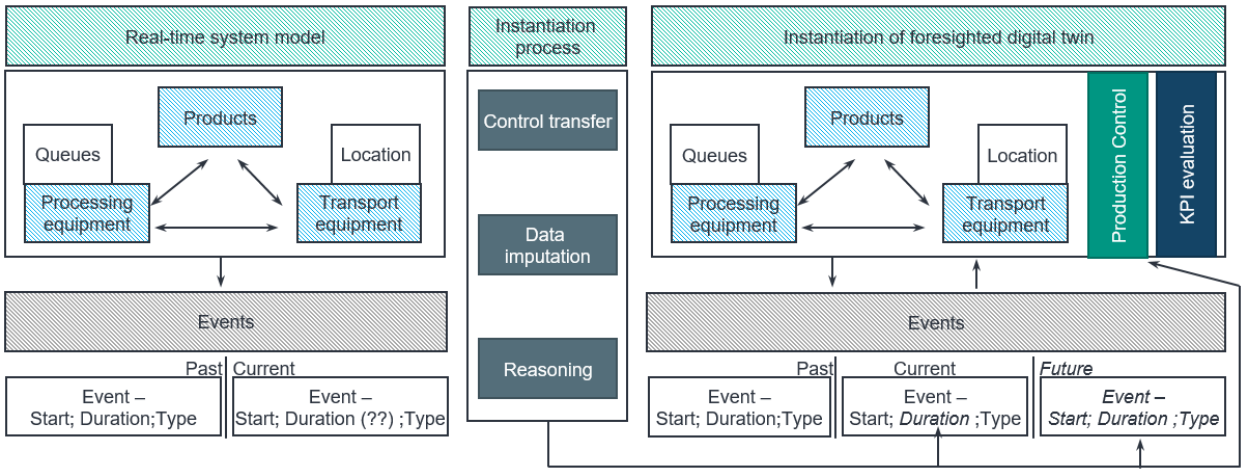


Figure 4: Digital twin architecture based on knowledge graph based DES and instantiation.

Constructing the digital representation makes use of fitting statistical distributions to observed past data. Within this study the detailed availability of real-time data from the fab's data based is not unlimited and, thus, constitutes the length of the observations regarded. Finding the perfect length of such a period is out of the scope of this study. In a similar vein past data can be mined to re-create the current status of lots, equipment and their interrelation provided sufficient data is available. The latter is used for the ex post evaluation so that the digital representation can serve as the up-to-the-minute model of the real system at any time.

Control transfer is not possible in form of interconnecting to the real control system in the course of this study. However, as only simple time constraints are regarded the evaluation time of one time constraint hardly exceeds a single day. Therefore, only short-term dispatching and scheduling have to be regarded. This can be even narrowed down as the evaluation task of the instantiated digital twin ends with a binary evaluation whether or not a time constraint within the simulation has been violated. Thus, the actual order in queues in positions after the lot in question would only be relevant in certain edge cases. This reduces the complexity and priority rules can be easily re-implemented and selected according to their actual fab usage at times. Missing data, in particular about current events, e.g. the duration and hence end of a current process for a certain lot, can imputed based on past data and the distributions mined. Due to the structure of a knowledge graph that corresponds to a pre-described ontology the model can be used for reasoning, which can identify missing links in new routings that have not been observed in the past, e.g. dedication of particular operations to certain machines.

## 4 TIME CONSTRAINT VIOLATION PREDICTION

To predict the violation of individual time constraints at the time the gate keeping decision of releasing or not releasing the time constraint lot is taken, the digital twin can be instantiated and the violation or adherence observed in the DES. The individual observation of a time constraint adherence however is insufficient to properly asses that violation probability (Lima et al. 2021). Thus, the digital twin is rolled out in a monte carlo style for multiple instances leading to an observation of the time constraint violations. The rollout strategy varies the seeds and scenarios obtained in the individual DES digital twin instances, to regard a more complete picture of the behavior as visualized in Figure 5. Multiple time constrained lots at the same gate keeping decision can similarly be evaluated. Production control decisions such as dispatching are taken by the transferred control logic as discussed in section 3, while later selection of time constrained lots are randomized to avoid an endless simulation of the proposed decision model in a simulation.

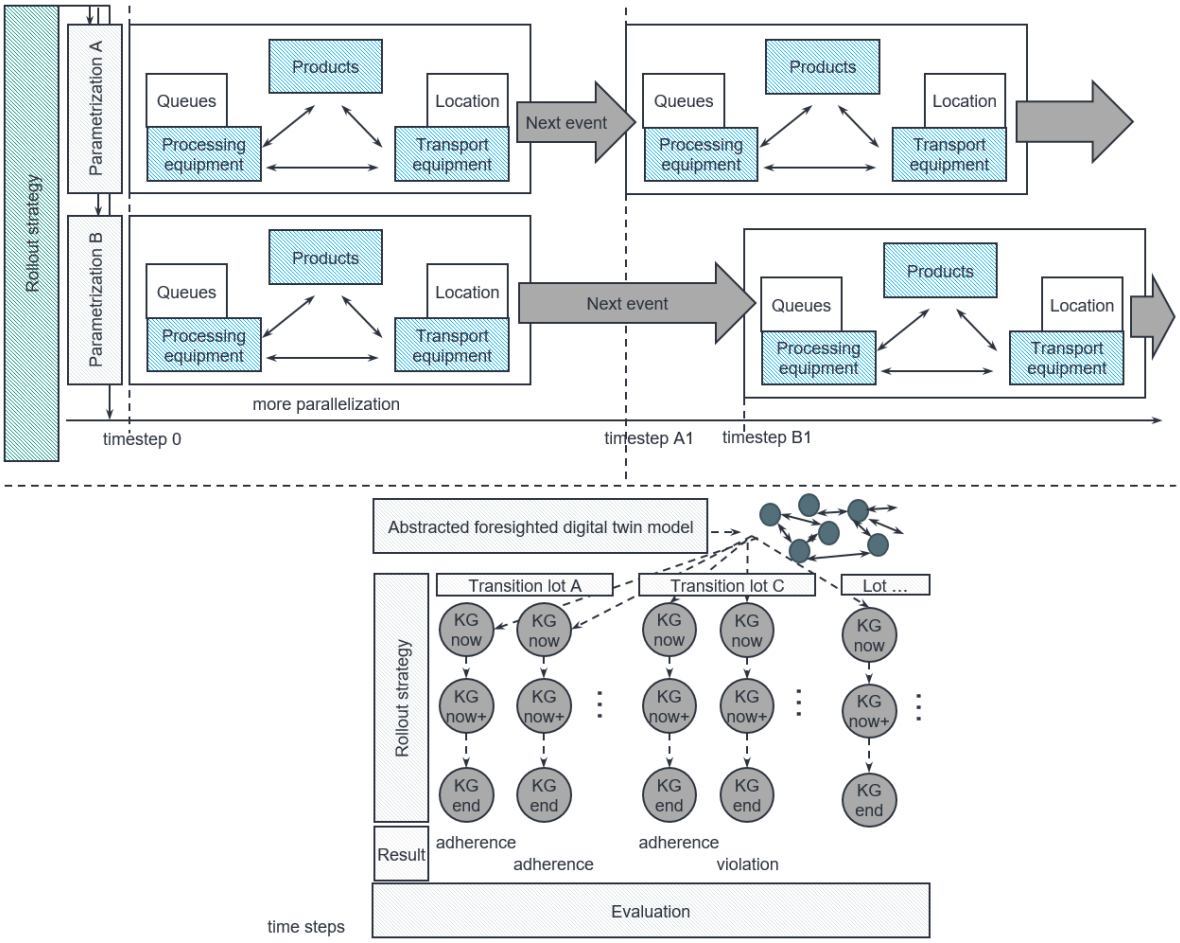


Figure 5: Rollout strategy and behavior.

For the evaluation of these rollouts a minimum regret policy of not releasing any time constrained lot that is at least violating the maximum time limit once or multiple times can be selected. Due to the interrelations within a semiconductor fab and the complex nature of their behavior (May et al. 2021) it is very likely that most time constrained lots will not satisfy this condition and an infeasible control logic is implemented. Thus, using a violation probability based approach (May et al. 2021) has advantages. The violation probability is obtained from the rollout sampling and only if the time constraint violation risk does not exceed a specified value the lot can be released as visualized in Figure 6. The transition time probability distribution is estimated based on the rollout samples observed. An acceptable risk limit of exceeding the time constraint can be ex post evaluated to conform to the fab operations strategy. Such a hyperparameter tuning can be achieved according to the actual trade-off between correctly and incorrectly withhold time constrained lots (May et al. 2021). This makes up the decisional rule in so far as time constrained lots can only be released if their time constraint violation risk does not exceed the selected threshold.

## 5 VALIDATION

The proposed approach is validated in a real-world semiconductor fab based on ex post fab data for several months. Given the transition time based evaluation the approach is restricted to simple time constraints and timelink areas as complex time constraints require a more in depth analysis.

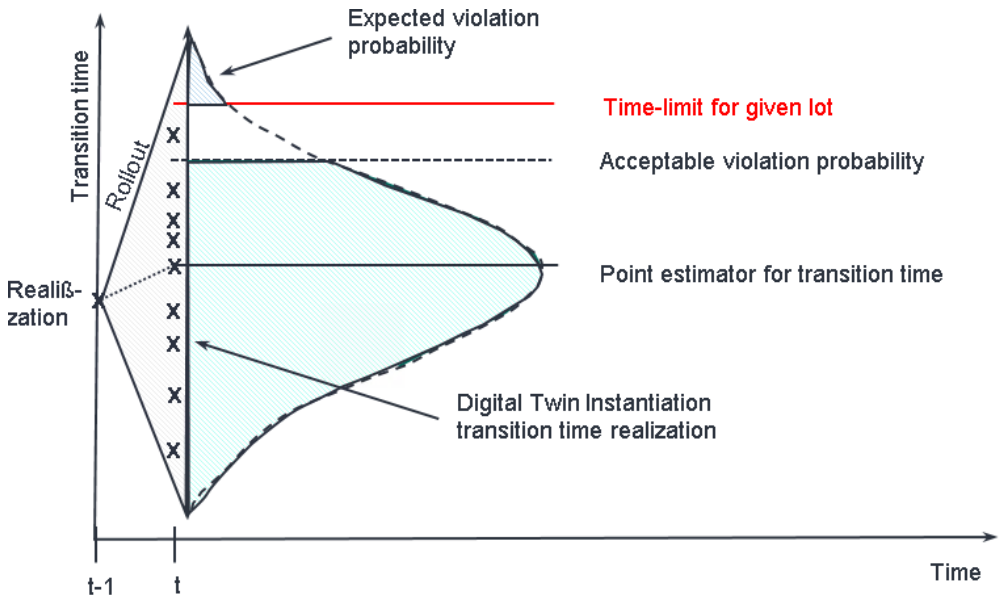


Figure 6: Evaluation of the time constraint violation probability and decisional rule.

Based on the ex post evaluation the model is compared to the real-world behavior in such a way that actual time constraint violations and adherences are re-evaluated and improvements over the state-of-the-art fab production control become visible following the approach of (May et al. 2024). Note that the dynamic nature of the decisional model leads to different decisions and results for an individual time constraint at different times so that withholding a lot at one decision does not mean it will be withhold for longer times. Table 1 shows the confusion matrix of the model's performance. More than two third of the actual violations, that would have occurred without any last minute interactions in the limited, regarded timeframe, could have been prevented with the proposed model showcasing the large improvement over the existing semiconductor fab control and manual interventions. However, a large number of lots is falsely predicted as a violation. Given the large, typically due to scrap, costs and environmental costs that come with any time constraint violation, the cost for short term falsely withholding a time constrained lot is almost negligible. Dominantly is that another, not time constrained lot is selected so that the throughput is not affected. In future work we will evaluate the effects on tardiness, what we clearly see in the validation is that a lot is not withhold for long as it is re-evaluated.

Table 1: Confusion matrix for the binary classification of time constraint violations with the proposed digital twin model

| | Actual violation | Actual adherence |
|---|---|---|
| Predicted violation | 21 | 681 |
| Predicted adherence | 10 | 2983 |

As a result the model can be positively validated as it indeed is capable of controlling the gate keeping decision of releasing time constrained lots in a semiconductor fab. It shows significant improvements over the state of the art in identifying violations, however, comes with a low precision. The low precision indicates a more conservative approach than the current control method, however, as proposed the evaluation of the rollout, here fine tuned with domain experts from the respective fab, can be adjusted. The decisional rule can be integrated in addition to the currently used method as additional measure by restricting the currents system choices with the proposed method. Thus, future work can improve both in the modeling and digital twin rollout as well as the actual decisional rule that is derived to improve precision.

## 6 CONCLUSION

The proposed approach to control the gate keeping of time constrained lots in semiconductor manfuacturing in a real-world setting and problem size is based on a digital twin that is instantiated multiple times to observe transition times and time constraint violations in the foresighted digital twin. Given the evaluation in this digital twin the actual release policy is constructed, based on a pre-selected maximum acceptable time constraint violation probability, where the probability is derived from the digital twin behavior observations. The model's real-world validation shows a reduction in time constraint violations of more than 67%, however comes at the cost of a large number of falsely predicted violations. The proposed methods validation is with ex post data so that real violations are re-evaluated ex ante to confirm the approaches efficacy. As a follow up the integration into the real fab is outstanding and to be reported in future papers. In Future work can improve the approach on the digital twin level, decision rule and possibility to include complex time constraints and evaluate their violation probability.

## ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with regard to project number LA 2351/88-1. The authors additionally would like to thank the Karlsruhe Institute of Technology (KIT) Academy for Responsible Research, Teaching, and Innovation (ARRTI) for supporting the research project "Artificial Intelligence for sustainable production planning (AI4NAPP)".

## REFERENCES

Altenmüller, T., T. Stüker, B. Waschneck, A. Kuhnle and G. Lanza. 2020. "Reinforcement learning for an intelligent and autonomous production control of complex job-shops under time constraints". *Production Engineering* 14:319–328 https://doi.org/10.1007/s11740-020-00967-8.

An, Y.-J., Y.-D. Kim, and S.-W. Choi. 2016. "Minimizing makespan in a two-machine flowshop with a limited waiting time constraint and sequence-dependent setup times". *Computers & Operations Research* 71:127–136 https://doi.org/10.1016/j.cor.2016.01.017.

Arima, S., A. Kobayashi, Y.-F. Wang, K. Sakurai and Y. Monma. 2015. "Optimization of Re-Entrant Hybrid Flows With Multiple Queue Time Constraints in Batch Processes of Semiconductor Manufacturing". *IEEE Transactions on Semiconductor Manufacturing* 28(4):528–544 https://doi.org/10.1109/TSM.2015.2478281.

Bixby, R., R. Burda, and D. Miller. 2006. "Short-Interval Detailed Production Scheduling in 300mm Semiconductor Manufacturing using Mixed Integer and Constraint Programming". In *The 17th Annual SEMI/IEEE ASMC 2006 Conference*, edited by N. Govind and J. Tyminski, 148–154. (Boston, USA, 22nd - 24th May. 2006) https://doi.org/10.1109/ASMC.2006.1638740.

Chien, C.-F., C.-J. Kuo, and C.-M. Yu. 2020. "Tool allocation to smooth work-in-process for cycle time reduction and an empirical study". *Annals of Operations Research* 290(1-2):1009–1033 https://doi.org/10.1007/s10479-018-3034-5.

Ciccullo, F., M. Pero, G. Pirovano, and A. Sianesi. 2014. "Scheduling batches with time constraints in a job shop system: developing two approaches for semiconductor industry". In *XIX Summer School "Francesco Turco" - Industrial Mechanical Plants*, edited by M. Bevilacqua, 12. (Senigalli, Italy, 9th - 12th Sep. 2014).

Hong, T.-Y., C.-F. Chien, and H.-P. Chen. 2023. "UNISON framework of system dynamics-based technology acquisition decision for semiconductor manufacturing and an empirical study". *Computers & Industrial Engineering*:109012 https://doi.org/10.1016/j.cie.2023.109012.

Kim, H.-J. and J.-H. Lee. 2017. "A Branch and Bound Algorithm for Three-Machine Flow Shop with Overlapping Waiting Time Constraints". *IFAC-PapersOnLine* 50(1):1101–1105 https://doi.org/10.1016/j.ifacol.2017.08.391.

Klemmt, A. and L. Mönch. 2012. "Scheduling jobs with time constraints between consecutive process steps in semiconductor manufacturing". In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, 1–10. (Berlin, Germany, 9th - 12th Dec. 2012) https://doi.org/10.1109/WSC.2012.6465235.

Kobayashi, A., T. Kuno, and S. Arima. 2013. "Re-entrant flow control in Q-time constraints processes for actual applications". In *2013 e-Manufacturing & Design Collaboration Symposium (eMDC)*, edited by L. Tuung, 1–4. (Hsinchu, Taiwan, 6th Sep. 2013) https://doi.org/10.1109/eMDC.2013.6756052.

Kopp, D., M. Hassoun, A. Kalir, and L. Mönch. 2020. "Integrating Critical Queue Time Constraints Into SMT2020 Simulation Models". In *Proceedings of the 2020 Winter Simulation Conference (WSC)*, 1813–1824. (Vienna, Austria, 14th - 18h Dec. 2020) https://doi.org/10.1109/WSC48552.2020.9383889.

Kuhnle, A., M. C. May, L. Schäfer, and G. Lanza. 2022. "Explainable reinforcement learning in production control of job shop manufacturing system". *International Journal of Production Research* 60(19):5812–5834 https://doi.org/10.1080/00207543.2021.1972179.

Lee, J.-Y. 2020. "A genetic algorithm for a two-machine flowshop with a limited waiting time constraint and sequence-dependent setup times". *Mathematical Problems in Engineering* 2020 https://doi.org/10.1155/2020/8833645.

Lima, A., V. Borodin, S. Dauzère-Pérès, and P. Vialletelle. 2019. "Sampling-based release control of multiple lots in time constraint tunnels". *Computers in Industry* 110:3–11 https://doi.org/10.1016/j.compind.2019.04.014.

Lima, A., V. Borodin, S. Dauzère-Pérès, and P. Vialletelle. 2017a. "Analyzing different dispatching policies for probability estimation in time constraint tunnels in semiconductor manufacturing". In *2017 Winter Simulation Conference (WSC)*, 3543–3554. (Las Vegas, NV, USA, 3th - 6th Dec. 2017) https://doi.org/10.1109/WSC.2017.8248068.

Lima, A., V. Borodin, S. Dauzère-Pérès, and P. Vialletelle. 2017b. "A decision support system for managing line stops of time constraint tunnels: FA, IE". In *2017 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, edited by R. Dover and D. LeCunff, 309–314. (Saratoga Springs, NY, USA, 15th - 18th May. 2017) https://doi.org/10.1109/ASMC.2017.7969250.

Lima, A., V. Borodin, S. Dauzère-Pérès, and P. Vialletelle. 2021. "A sampling-based approach for managing lot release in time constraint tunnels in semiconductor manufacturing". *International Journal of Production Research* 59(3):860–884 https://doi.org/10.1080/00207543.2020.1711984.

Maleck, C. and T. Eckert. 2017. "A comparison of control methods for production areas with time constraints and tool interruptions in semiconductor manufacturing". In *2017 40th International Spring Seminar on Electronics Technology (ISSE)*, edited by A. Hamacek and N. Hinov, 1–6. (Sofia, Bulgaria, 10th - 14th May. 2017) https://doi.org/10.1109/ISSE.2017.8000944.

Maleck, C., G. Nieke, K. Bock, D. Pabst, M. Schulze and M. Stehli. 2019. "A Robust Multi-Stage Scheduling Approach for Semiconductor Manufacturing Production Areas with Time Contraints". In *2019 30th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, edited by A. Jain and F. Levitov, 1–6. (Saratoga Springs, NY, USA, 6th - 9th May. 2019) https://doi.org/10.1109/ASMC.2019.8791779.

Maleck, C., G. Nieke, K. Bock, D. Pabst and M. Stehli. 2018. "A comparison of an CP and MIP approach for scheduling jobs in production areas with time constraints and uncertainties". In *2018 Winter Simulation Conference (WSC)*, 3526–3537. (Gothenburg, Sweden, 9th - 12th Dec. 2018) https://doi.org/10.1109/WSC.2018.8632404.

May, M. C., A. Albers, M. D. Fischer, F. Mayerhofer, L. Schäfer and G. Lanza. 2021. "Queue Length Forecasting in Complex Manufacturing Job Shops". *Forecasting* 3(2):322–338 https://doi.org/10.3390/forecast3020021.

May, M. C., L. Behnen, A. Holzer, A. Kuhnle and G. Lanza. 2021. "Multi-variate time-series for time constraint adherence prediction in complex job shops". *Procedia CIRP* 103:55–60 https://doi.org/10.1016/j.procir.2021.10.008.

May, M. C., L. Kiefer, A. Kuhnle, and G. Lanza. 2022. "Ontology-Based Production Simulation with OntologySim". *Applied Sciences* 12(3) https://doi.org/10.3390/app12031608.

May, M. C., S. Maucher, A. Holzer, A. Kuhnle and G. Lanza. 2021. "Data analytics for time constraint adherence prediction in a semiconductor manufacturing use-case". *Procedia CIRP* 100:49–54 https://doi.org/10.1016/j.procir.2021.05.008.

May, M. C., J. Neidhöfer, T. Körner, L. Schäfer and G. Lanza. 2022. "Applying natural language processing in manufacturing". *Procedia CIRP* 115:184–189 https://doi.org/10.1016/j.procir.2022.10.071.

May, M. C., C. Nestroy, L. Overbeck, and G. Lanza. 2024. "Automated model generation framework for material flow simulations of production systems". *International Journal of Production Research* 62(1-2):141–156.

May, M. C., J. Oberst, and G. Lanza. 2024. "Managing product-inherent constraints with artificial intelligence: production control for time constraints in semiconductor manufacturing". *Journal of Intelligent Manufacturing*:1–18 https://doi.org/10.1007/s10845-024-02472-6.

May, M. C., L. Overbeck, M. Wurster, A. Kuhnle and G. Lanza. 2021. "Foresighted digital twin for situational agent selection in production control". *Procedia CIRP* 99:27–32 https://doi.org/10.1016/j.procir.2021.03.005.

Min, Q., Y. Lu, Z. Liu, C. Su and B. Wang. 2019. "Machine learning based digital twin framework for production optimization in petrochemical industry". *International Journal of Information Management* 49:502–519 https://doi.org/10.1016/j.ijinfomgt.2019.05.020.

Mönch, L., J. W. Fowler, S. Dauzère-Pérès, S. J. Mason and O. Rose. 2011. "A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations". *Journal of Scheduling* 14(6):583–599 https://doi.org/10.1007/s10951-010-0222-9.

Mönch, L., J. W. Fowler, and S. J. Mason. 2013. *Production planning and control for semiconductor wafer fabrication facilities: Modeling, analysis, and systems*. 52 ed.
Operations Research / Computer Science Interfaces Series. New York, NY: Springer https://doi.org/10.1007/978-1-4614-4472-5.

Nattaf, M., S. Dauzère-Pérès, C. Yugma, and C.-H. Wu. 2019. "Parallel machine scheduling with time constraints on machine qualifications". *Computers & Operations Research* 107:61–76 https://doi.org/10.1016/j.cor.2019.03.004.

Negri, E., L. Fumagalli, and M. Macchi. 2017. "A review of the roles of digital twin in CPS-based production systems". *Procedia Manufacturing* 11:939–948 https://doi.org/10.1016/j.promfg.2017.07.198.

Ono, A., S. Kitamura, and K. Mori. 2006. "Risk Based Capacity Planning Method for Semiconductor Fab with Queue Time Constraints". In *2006 IEEE International Symposium on Semiconductor Manufacturing*, edited by H. Sasaki and T. Ohmi, 49–52. (Tokyo, Japan, 25th - 27st Sep. 2006) https://doi.org/10.1109/ISSM.2006.4493020.

Pappert, F. S., T. Zhang, O. Rose, F. Suhrke, J. Mager and T. Frey. 2016. "Impact of time bound constraints and batching on metallization in an opto-semiconductor fab". In *Proceedings of the 2016 Winter Simulation Conference (WSC)*, 2947–2957. (Washington D.C., USA, 11th - 14th Dec. 2016) https://doi.org/10.1109/WSC.2016.7822329.

Pirovano, G., F. Ciccullo, M. Pero, and T. Rossi. 2020. "Scheduling batches with time constraints in wafer fabrication". *International Journal of Operational Research* 37(1):1–31 https://doi.org/10.1504/IJOR.2020.104222.

Robinson, J. K. and R. Giglio. 1999. "Capacity planning for semiconductor wafer fabrication with time constraints between operations". In *Proceedings of the 31st Conference on Winter Simulation: Simulation—a bridge to the future*, Volume 1, 880–887. (Phoenix, AZ, USA, 5th - 8th Dec. 1999) https://doi.org/10.1145/324138.324545.

Sadeghi, R., S. Dauzère-Pérès, C. Yugma, and G. Lepelletier. 2015. "Production control in semiconductor manufacturing with time constraints". In *2015 26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, edited by C. Ordonio and J. Schaller, 29–33. (Saratoga Springs, NY, USA, 3rd - 6th May. 2015) https://doi.org/10.1109/ASMC.2015.7164446.

Scholl, W. and J. Domaschke. 2000. "Implementation of modeling and simulation in semi- conductor wafer fabrication with time constraints between wet etch and furnace operations". *IEEE Transactions on Semiconductor Manufacturing* 13(3):273–277 https://doi.org/10.1109/66.857935.

Toyoshima, N., T. Hasegawa, K. Wu, and S. Arima. 2013. "Proactive control of engineering operations and lot loadings of product-mix and re-entrant in Q-time constraints processes". In *2013 e-Manufacturing & Design Collaboration Symposium (eMDC)*, edited by L. Tuung, 1–4. (Hsinchu, Taiwan, 6th Sep. 2013) https://doi.org/10.1109/eMDC.2013.6756046.

Tu, Y.-M. and C.-L. Chen. 2011. "Model to determine the capacity of wafer fabrications for batch-serial processes with time constraints". *International Journal of Production Research* 49(10):2907–2923 https://doi.org/10.1080/00207541003730854.

Tu, Y.-M. and H.-N. Chen. 2009. "Tool portfolio planning in the back-end process of wafer fabrication with sequential time constraints". *Journal of the Chinese Institute of Industrial Engineers* 26(1):60–69 https://doi.org/10.1080/10170660909509122.

Uçar, E., M.-A. Le Dain, and I. Joly. 2020. "Digital technologies in circular economy transition: evidence from case studies". *Procedia CIRP* 90:133–136 https://doi.org/10.1016/j.procir.2020.01.058.

Uhlemann, T. H.-J., C. Lehmann, and R. Steinhilper. 2017. "The digital twin: Realizing the cyber-physical production system for industry 4.0". *Procedia CIRP* 61:335–340 https://doi.org/10.1016/j.procir.2016.11.152.

Valet, A., T. Altenmüller, B. Waschneck, M. C. May, A. Kuhnle and G. Lanza. 2022. "Opportunistic maintenance scheduling with deep reinforcement learning". *Journal of Manufacturing Systems* 64:518–534 https://doi.org/10.1016/j.jmsy.2022.07.016.

Wang, M., S. Srivathsan, E. Huang, and K. Wu. 2018. "Job Dispatch Control for Production Lines With Overlapped Time Window Constraints". *IEEE Transactions on Semiconductor Manufacturing* 31(2):206–214 https://doi.org/10.1109/TSM.2018.2826530.

Winkler, T., P. Barthel, and R. Sprenger. 2016. "Modeling of complex decision making using forward simulation". In *2016 Winter Simulation Conference (WSC)*, 2982–2991. IEEE.

Wu, C.-H., W.-C. Chien, Y.-T. Chuang, and Y.-C. Cheng. 2016. "Multiple product admission control in semiconductor manufacturing systems with process queue time (PQT) constraints". *Computers & Industrial Engineering* 99:347–363 https://doi.org/10.1016/j.cie.2016.04.003.

Yu, T.-S., H.-J. Kim, C. Jung, and T.-E. Lee. 2013. "Two-stage lot scheduling with waiting time constraints and due dates". In *Proceedings of the 2013 Winter Simulations Conference (WSC)*, 3630–3641. (Washington D.C., USA, 8th - 13th Dec. 2013) https://doi.org/10.1109/WSC.2013.6721724.

Zhang, T., F. S. Pappert, and O. Rose. 2016. "Time bound control in a stochastic dynamic wafer fab". In *Proceedings of the 2016 Winter Simulation Conference (WSC)*, 2903–2911. (Washington D.C., USA, 11th - 14th Dec. 2016) https://doi.org/10.1109/WSC.2016.7822325.

Zhou, L., C. Lin, B. Hu, and Z. Cao. 2019. "A Cuckoo Search-Based Scheduling Algorithm for a Semiconductor Production Line with Constrained Waiting Time". In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, edited by W. Shen, J. Li, and A. Matta, 338–343. (Vancouver, Canada, 22nd - 26th Aug. 2019) https://doi.org/10.1109/COASE.2019.8842869.

Zhou, Y. and K. Wu. 2017. "Heuristic simulated annealing approach for diffusion scheduling in a semiconductor Fab". In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, edited by R. Lee, X. Cui, and S. Xu, 785–789. (Wuhan, China, 24th - 26th May. 2017) https://doi.org/10.1109/ICIS.2017.7960099.

Ziarnetzky, T., L. Mönch, T. Ponsignon, and H. Ehm. 2017. "Rolling horizon planning with engineering activities in semiconductor supply chains". In *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*, edited by X. Guan and Q. Zhao, 1024–1025. (Xi'an, China, 20th - 23rd Aug. 2017): IEEE https://doi.org/10.1109/COASE.2017.8256237.

## AUTHOR BIOGRAPHIES

**MARVIN CARL MAY** is Chief Engineer and postdoctoral researcher at the wbk Institute of Production Science at the Karlsruhe Institute of Technology. His research interests include Production Planning and Control, Product-Production-CoDesign, Simulation based optimization and Machine Learning for optimal control within manufacturing and an emphasize on semiconductor manufacturing environments. His email address is marvin.may@kit.edu.

**LARS KIEFER** is a researcher at the wbk Institute of Production Science at the Karlsruhe Institute of Technology. His research interests include Production Planning and Control, Simulations with Ontology integration and Machine Learning for control within manufacturing. His email address is lars.kiefer@alumni.kit.edu.

**GISELA LANZA** is a full professor and member of the management board at the Institute of Production Science (wbk) of the Karlsruhe Institute of Technology (KIT). She heads the Production Systems division dealing with the topics of global production strategies, production system planning, and quality assurance in research and industrial practice. Her email address is gisela.lanza@kit.edu.