Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

IMPACT OF TIME BOUND CONSTRAINTS AND BATCHING ON METALLIZATION IN AN OPTO-SEMICONDUCTOR FAB

Falk Stefan Pappert
Tao Zhang
Oliver Rose

Department of Computer Science
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
Neubiberg, 85577, GERMANY

Fabian Suhrke
Jonas Mager
Thomas Frey

Osram Opto Semiconductors GmbH
Leibnizstraße 4
Regensburg, 93055, GERMANY

ABSTRACT

Time bound sequences are constraints deemed necessary to ensure product quality and avoid yield loss due to time dependent effects. Although they are commonly applied in production system control they cause severe logistical challenges. In this paper, we evaluate the effects of time constraints in combination with batching on a real metallization work center of an opto-semiconductor fab. We use simulation to analyze the impact of these production constraints and point out potentials to increase work center performance. We have a closer look at the required planning horizon, the influence of dedication, the capacity loss due to time bounds and the effects of batching strategies on wafer cost. Our results show the importance to tackle these issues. Furthermore, we will discuss actions taken in response to the experiments.

1 INTRODUCTION

Time bound sequences are a common constraint in semiconductor manufacturing. These constraints represent time bounds in which a number of succeeding process steps should be performed. Violating these constraints usually necessitates rework or, even worse, the scrapping of the affected wafers. The reason for these constraints is usually to keep particle contamination and surface reactions to a level where it does not influence the process quality. To avoid violating these time bounds effective dispatching or scheduling is used to keep them to a statistical minimum. The effect on the system is usually that some lots, batches or wafers are on hold until sufficient resources are available to ensure that they can be processed before the time runs out. While trying to ensure non-violation release strategies we basically trade equipment utilization for cycle time. There are several approaches in literature to tackle this issue. Robinson (1998) and Robinson and Giglio (1999) presented a basic approach to capacity planning with time bound constraints and calculations to estimate time bound violations. Scholl and Domaschke (2010) proposed a Kanban-type approach where tool capacity is limited directly. Klemmt and Mönch (2012) proposed a heuristic and an MIP based approach of scheduling lots in a time bound sequence. In general, there is always some loss of capacity to ensure as few as possible violations.

In this paper, we want to present a first study of a coating work center and its time bound sequences to evaluate the impact of different logistical characteristics on the system. The system at hand is a group of batch coating equipment with a number of tools supporting preprocessing and handling steps. The time bounds considered are a mix of time bound sequences with and without intermediate steps using a number

of different time targets. The system is modeled from a productive manufacturing line of an opto-semiconductor manufacturer which in contrast to traditional semiconductor manufacturers faces a broader spectrum of materials used for coating. This, in turn, increases the number of different recipes, tool dedications and processes in the coating workshop and therefore its complexity.

In this paper, we will first present the real system and its environment and give an overview on the model characteristics. Then, we will discuss several experiments and determine the influence of logistical characteristics on the system.

2 THE MODEL

In this section, we will introduce the simulation environment and the model used for our experiments.

2.1 Model Environment

With the recent rise in demand for LED, Osram Opto Semiconductor is drastically increasing its efforts to collect and use fab data to improve their logistic processes. Although these efforts have come far, data availability has to be further improved to compete with leaders from traditional semiconductor fields. Therefore, the completely automated generation of a workshop model is not feasible at the moment. At the same time, a completely manual approach to model building is similarly unfeasible as the amount of information to create a simulation model for a specific workshop is near to impossible to maintain in an environment of permanently changing products, product mix and tool set. Hence, we use a semi-automated approach to generate models. The approach is visualized in Figure 1. Furthermore, considering increased involvement of users, an Excel frontend was chosen as a familiar user interface to increase acceptance and compatibility with input data not yet available in data bases. In our concept, we provide the user with an Excel template which is able to import raw fab data from a number of data bases. In a second step, this data can be adapted and missing data is added to create a full dataset. This is done in an Excel spreadsheet to provide the users with their daily work environment to reduce usability issues and training time. From this user-interaction based format we generate a second Excel file translating all information into a machine readable format which is compatible with our simulation meta model. During this first transformation we make the model machine readable and transform values and measures from the units as they are used on a day-by-day basis to a set of systematic units to standardize data.

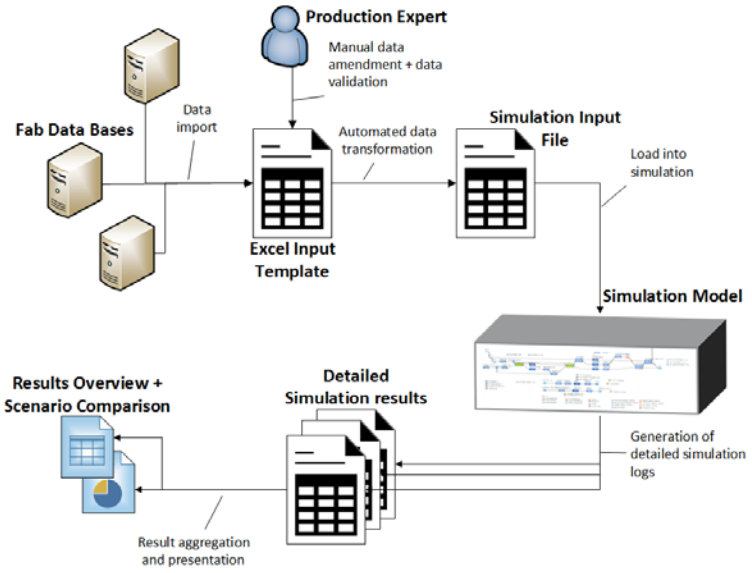

Figure 1: Simulation system overview.

We decided to use this second step to improve transparency for the model transformation process and to help to find transformation problems which might occur because of manual data entries which usually are more prone to errors than automatically generated data. The resulting machine readable Excel model is read by a generic AnyLogic model and then creates and configures tool groups, routes and the controller.

The implemented simulation model is basically a lot generator, a data structure for tool groups and a sink. During initialization, tool groups are created and parameterized. They are connected to a central controller which handles all scheduling and dispatching decisions. Therefore, it is quite easy to replace scheduling and dispatching approaches and try out new ideas on the same system. Simulation runs usually only take a couple of seconds, the exact time mainly depends on load levels and queue sizes. Most of the simulation time is actually used by the implemented logistical strategies, e.g., dispatching rules, which take longer for longer queues. More details on the controller and system architecture of the simulation can be found in Zhang et al. (2016).

After every simulation run, the results are written to a raw data output Excel file. These files include detailed information on every lot's movement through the system as well as detailed logs on tool states and basic statistics. The raw data files of all replications of all design points of an experiment are then concentrated into a single result overview file. We decided to use a two level approach here as well, the main reason to do this is customizability for new concepts and comfortability when evaluating simulation results. During the development of this simulation environment and in its future use it was used and will be used by users and developers with very different skill sets. On one hand, most users will have no problem with using spreadsheet calculations or smaller VBA scripts to calculate key performance indicators. Thus, they will be able to create and use their own analyses with the help of raw data files. On the other hand, we cannot expect users to be able to read and modify complex source code in Java. Therefore, we consider it to be worthwhile to have a two layer approach which is comfortable to use for user groups and only on rare occasions needs additional support from specialized software developers.

2.2 The Real World

The workshop we focus on in our experiments is part of the deposition processes. In deposition processes, wafers are coated with materials, usually metals or metal alloys, in our case Physical Vapor Deposition (PVD) is used. Wafers are processed in batches which are mounted onto a carrier. In a vacuum, the deposition material is heated until it evaporates. The particles in the process chamber then start to settle on the carriers and wafers in the process chamber and create a thin coating. Depending on the material and the target thickness of the coating, this processes usually takes somewhere from 45 minutes up to 8 hours.

Deposition is a very common area in semiconductor manufacturing to enforce time bounds as wafers need to be prepared for this process by one or more cleaning and handling steps. After preparing the surface of the wafer for the deposition process the surface immediately starts to deteriorate again. This deterioration is commonly caused by particles settling over time as well as oxidation reactions of the surface. Depending on the process, negative effects of deterioration (i.e., oxidation) reduce the quality of the product and yield after 30 minutes to 48 hours to an extend where it is no longer economically feasible to start the deposition process. In case the surface deteriorated too far, another round of preparation steps may be needed to re-prepare the wafers for deposition. While it is usually possible to repeat these preparation steps there is a maximum number of retries where each one is wasting capacity of tools involved in the process. In some rare cases, it might not even be possible at all causing expensive wafers to be turned into scrap immediately. Scheduling and dispatching for these time bound sequences usually trade in a certain amount of capacity for a degree of safety when controlling material flow in these areas.

2.3 The Model

As mentioned above the model we used in this study represents a part of the deposition tool groups in an opto-semiconductor factory. We consider 39 representative routes through the system representing different layers and product groups. Most of the routes have time constraints between 30 minutes and 48 hours. There are on average about 420 lot starts per week with huge variation in lot sizes depending on the product. Due to breakage in previous process steps and single wafers being on hold for process control the actual number of wafer in a lot tends to be slightly lower than the maximum. For the experiments here, we used planned production starts for each week according to fab data and reduced wafer numbers in each lot with the help of an expected loss distribution.

The simulation model represents a total of 58 tools. Deposition tools marking the end of the time bound sequence. The other tools represent the 10 groups of cleaning and preparation tools. All tools in the studied system are batch tools supporting different batch sizes depending on the processes and wafer sizes. The deposition tools have, in addition to breakdowns, detailed maintenance schedules to cater for weekly mandatory maintenance and regular material top-up. Furthermore, we have to consider tool dedication as a major limiting factor for capacity and throughput. The 26 different processes on the deposition tools are qualified only on a subset of these deposition tools making planning more difficult and breakdowns harder to cope with.

3 EXPERIMENTS

The main focus and reason for the following experiments is to improve the knowledge about the system and to understand and evaluate the effects of time bounds on the system at hand. It is common knowledge that time bound sequences can have a significant influence on systems but the magnitude of this influence depends on the system itself. To showcase the results we will mainly present operating curves as they provide a good impression of the effects on capacity and cycle time. All design points were simulated at least 20 times until we obtained statistically significant results. We will first have a look at the tradeoff between material flow and product cost as a result of different batching strategies. We will then move on to analyses of dedication and planning horizons. Finally, we will have a look at the general loss of capacity caused by time bounds in the system.

3.1 Material Flow vs Material Cost

A common challenge with batch processes is to decide which batch rules to implement, especially the completeness of a batch which is to be enforced. There is generally a tradeoff between the time an uncompleted batch is waiting for further wafers and the material and process cost for the step. On one hand, the fuller a processed batch is, the better the material efficiency for this process. On the other hand, the longer an uncompleted batch waits for further wafers the higher the cycle times for all wafers waiting will be. Therefore, we have a hard time to quantify tradeoff between material and logistical cost. This challenge is especially interesting as the process mix includes very different product quantities for different products. There is usually no issue to wait a moment for high runners, but for low volume products this could result in unacceptable high flow factors.
Batch strategies are defined by three basic parameters.

- Minimum batch size (minBatch) – Is the minimal number of wafers a batch should have before it is started.
- Maximum batch size (maxBatch) – represents the maximal number of wafers a batch is allowed to contain. This is usually based on tool or carrier capacity but may also be for reasons of process stability.

- Maximum waiting time (maxWait) – represents the maximal time an uncompleted batch should wait for further wafers before it is allowed to start the process even though it has not yet reached minBatch

In this experiment, we compared three general setups.

- Full batch – this simply means minBatch is set to the same value as maxBatch. Therefore only full batches and batches violating maxWait are allowed to start.
- 66%/100% - This strategy is based on the assumption, that high runners usually don't have to wait for long before they complete a full batch. Hence, they are kept at minBatch equal to maxBatch. On the flip side low runners are assigned minBatch values as low as 66% of their maxBatch value. The fullest batch is preferred.
- The current Osram strategy- Is basically similar to 66%/100% but with the minBatch setting put closer to 90% with slight variations based on product priority.

The results are shown in Figure 2 and Figure 3.

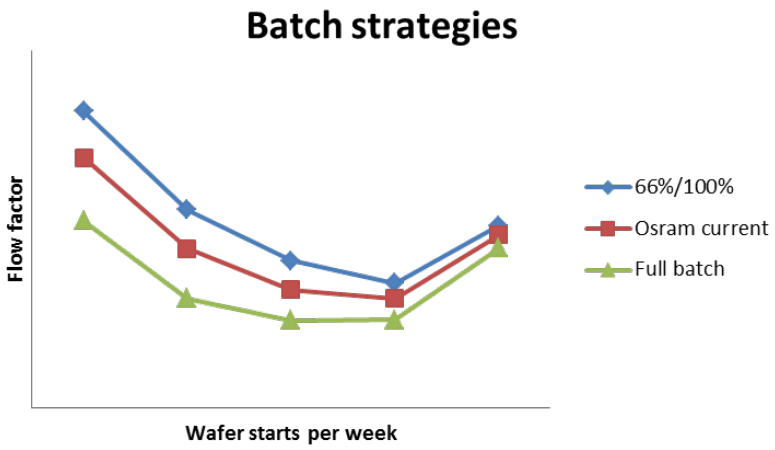

Figure 2: Comparison of batch strategies and their effect on flow factors.

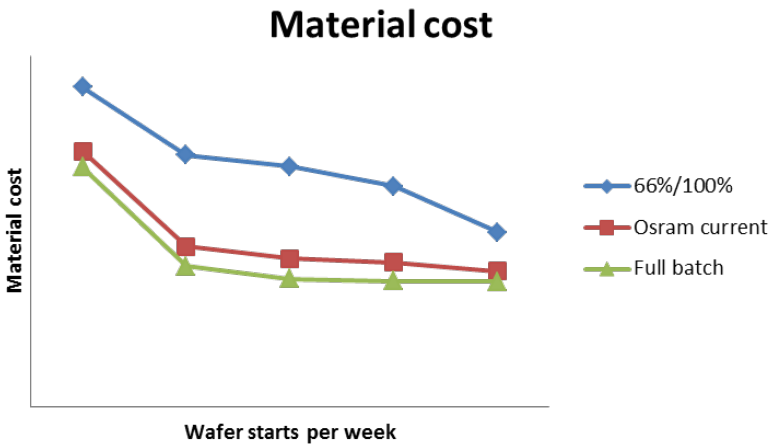

Figure 3: Comparison of batch strategies and their effect on cost per wafer.

Figure 2 shows the expected influence of batch strategies on cycle times. The fuller a batch needs to be, the higher the waiting and therefore cycle times of products. Starting from very low volume and therefore very high flow factors (cycle time divided by raw processing time) the influence of batch building diminishes the higher the work center is utilized. Once the fab loading comes closer to 100% the flow factor rises again.

In Figure 3, we see the effects of these strategies on wafer cost. Although we cannot disclose detailed cost information here, it is quite apparent that giving a little leeway with minBatch results in only small cost increases while giving to much can result in a drastic jump in cost. The reason for this could be found in low runners having higher material cost.

In this study, we showed that the current batching strategy represents an effective working point and we will therefore implement this strategy into an upcoming dispatcher ruleset.

3.2 Planning Horizon

In the traditional semiconductor industry, complex dispatching systems and manufacturing execution systems are quite common. They are able to schedule every lot based on detailed information tracked and reported from all over the manufacturing floor. In the opto-semiconductor area, this is still on the implementation agenda. Currently operators are still a major resource for transport and processing that have a significant influence on how well material is moved through the system. In contrast to complex scheduling systems, it is challenging for operators to plan numerous jobs ahead while ensuring machines are not idle and time bounds are not violated. According to experts from the shopfloor it is a reasonable assumption for operators to being able to manually keep track of 1 to 2 batches in the time bound sequence for each main tool. One batch basically means, only once the main tool is free again a new batch will be started into the preparation, while two usually means there is one batch in the main tool while another one is in preparation. One and two sound to be very low values, but as every operator usually operates several tools this actually already involves quite some planning and includes a lot of communication with other operators. In this experiment, we refer to the planning horizon as the maximal number of batches scheduled for every single main tool. Therefore this value also represents the upper limit of batches in the time bound sequence for each main tool but does not cause batches to be released unless the dispatcher was able to schedule them without violating the time constraint.

In this experiment, we want to have a look at the influence of the planning horizon on the system performance and then evaluate the potential for improvement when operators get more support from a more elaborate dispatching system assuming reasonable compliance with it.

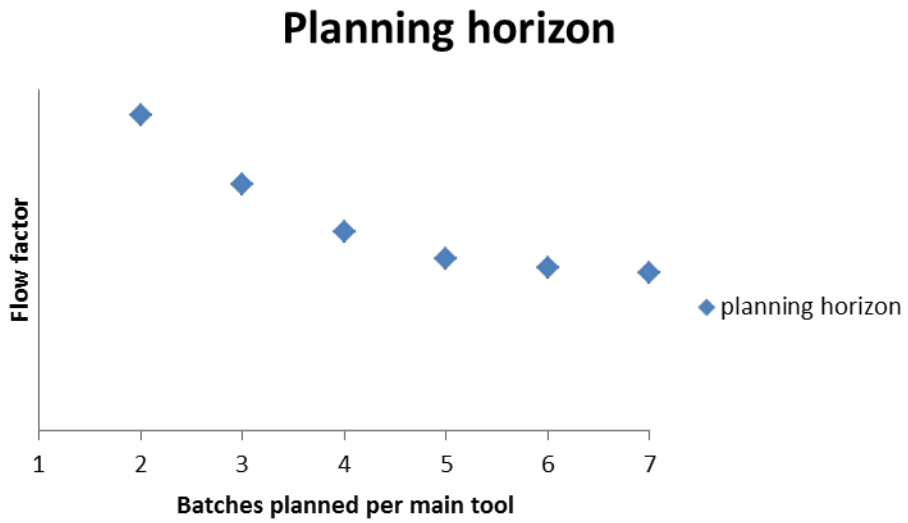

Figure 4: Influence of planning horizon on the systems performance.

Figure 4 shows an improvement with increasing planning horizon as expected. The influence in the system at hand is actually rather significant. A simple increase in batches scheduled can drastically improve flow factors without the need for additional process capacity.

Dispatching of lots with time bounds is currently rather inefficient because operators must handle the complex scheduling problem without software support. This leads to an increased risk that the operations do not achieve the planned capacity. By means of simulation, we have shown that an efficient time bound control system is necessary, which can reduce the flow factor by up to 25%. For this reason, we prioritized the implementation of time bound control within the real-time dispatcher.

3.3 Dedication

Tool dedication is a common characteristic in semiconductor industries. This constraint is usually caused by the need to qualify or configure tools in a way that they are able to handle a certain process. In job shops where tools are used for different processes it is often not possible to qualify all tools of a kind for all processes. This is in the simplest case caused by different setup requirements of different processes which outright prohibit complete qualification. In less restrictive situations, full qualification is often not economical for reasons of process stability and qualification effort. Usually with an increasing number of qualifications on a tool the effort increases to allow for all previous qualifications to stay valid. Furthermore, regular qualification checks are necessary to ensure product quality which is again more effort with an increasing number of qualifications.

In this experiment, we have a look at the effects of increasing the number of qualified tools for each process slightly to determine whether benefits of higher flexibility would outweigh the cost for additional qualifications. The additional qualifications are done on tools where this would not collide with current qualifications.

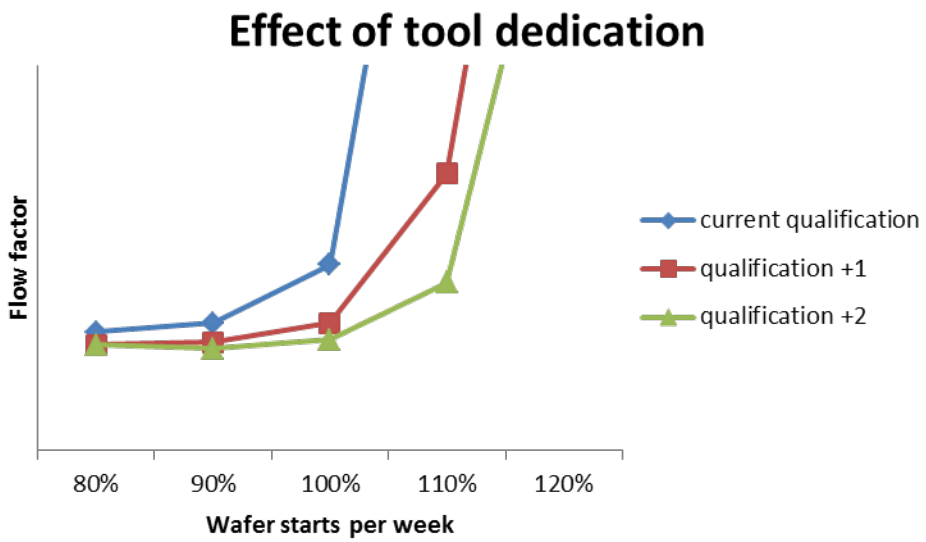

Figure 5: Influence of more broader qualification to the flow factor.

Figure 5 shows the improvement in the average flow factor with increasing qualification. The modeled system would be able to handle 10% more material and keeping its current average flow factor when increasing the number of qualified tools for each process.

These results show how essential a broad equipment qualification for new products is. The product engineers should be aware, especially for bottleneck equipment, to re-qualify short-term locked processes.

3.4 Capacity Cost of Time Bounds

We mentioned in the introduction and reiterated several times since then that time bound sequences waste capacity of the tools involved. The reasons for introducing time bounds are often valid and more often than not they are introduced to ensure product quality after product measurements indicated issues with the process. Still the major question remains how much capacity is actually wasted with time bound sequences.

In this experiment, we compare the operating curves of the system in its current state with the system without any time bounds.

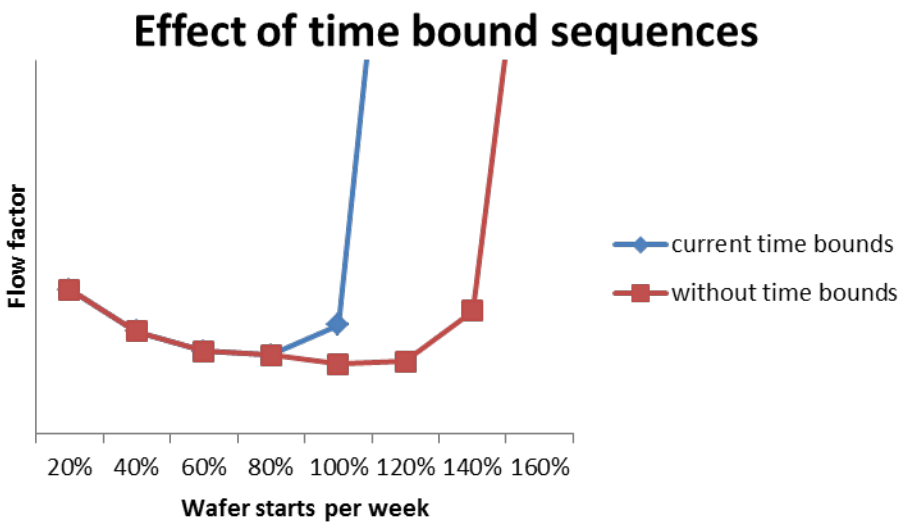

Figure 6: Effect of time bound sequences on flow factor and system capacity respectively.

Figure 6 shows the results of the experiment. As one would expect the system without any time bounds sequences could handle more material with much less effect on the flow factor than the system with the real world settings. Even knowing of a loss in capacity, it was not clear just how much capacity is lost due to time bound sequences. Of course, we realize that it is not possible to remove all time bounds. We see the area between both curves as an area of possible improvements. Reducing the number of time bounds or simply increasing the time given to each batch within the time bound offers a significant opportunity to reclaim capacity from the system.

With support of the simulation, the influence of time bounds on the material flow can be determined. Therefore, we are able to effectively motivate a review and rework of time bounds.

4 CONCLUSION

In this paper, we presented the model of a real coating work center in an opto-semiconductor fab. We evaluated the influence of several important factors of control and capacity on the system and were able to show the significance of time bound sequences and the need to control them. Especially, the loss of capacity due to time bound sequences demands a closer look on the system and further investigation. We hope that with the experiments presented here we can encourage other companies and researchers to consider the introduction of time bound constraints more carefully and reevaluate the cost of existing ones.

5 FUTURE RESEARCH

After this first set of general studies to evaluate the potential of the reviewed time bounds, we will focus on how to unleash the found potential in cooperation with process engineers. We see further research opportunities in methods of gradually removing time bounds from systems.

ACKNOWLEDGEMENTS

We would like to thank Twan van der Borgh for his help with the implementation of the simulation system.

REFERENCES

Klemmt, A., and Lars Mönch. 2012. "Scheduling Jobs with Time Constraints Between Consecutive Process Steps in Semiconductor Manufacturing." In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose and A.M. Uhrmacher, 1-10. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. .

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 Manufactoring Operations" *Journal of Scheduling* 14:583-599. Springer Science+Business Media

Robinson, J. K. 1998. "Capacity Planning in a Semiconductor Wafer Fabrication Facility with Time Constraints Between Process Steps." Ph.D. thesis, Department of Mechanical & Industrial Engineering, University of Massachusetts, Amherst, Massachusetts.

Robinson, J. K., and R. Giglio. 1999. "Capacity Planning for Semiconductor Wafer Fabrication with Time Constraints Between Operations." In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 880-887. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. .

Scholl, W., and J. Domaschke. 2000. "Implementation of Modeling and Simulation in Semiconductor Wafer Fabrication with Time Constraints Between Wet Etch and Furnace Operations." *IEEE Transactions on Semiconductor Manufacturing* 13 (3):273-277.

Zhang, T., F. S. Pappert, O. Rose. 2016. "Time Bound Control in a Stochastic Dynamic Wafer Fab" Submitted to *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. .

AUTHOR BIOGRAPHIES

FALK STEFAN PAPPERT is a Research Assistant and PhD student at Universität der Bundeswehr München as a member of the scientific staff of Prof. Dr. Oliver Rose at the Chair of Modeling and Simulation. His focus is on conceptual modelling approaches to simulation-based scheduling and optimization of production systems. He has received his M.S. degree in Computer Science from Dresden University of Technology. He is a member of GI. His email address is falk.pappert@unibw.de.

TAO ZHANG is a Ph.D. student working on production planning and scheduling at the Department of Computer Science of the Universität der Bundeswehr München, Germany. From 2007 to 2009 he received his Master in metallurgical engineering with the subject of production planning and scheduling in iron and steel industry from Chongqing University, China. He is involved in modeling and simulation of complex system and intelligent optimization algorithms. His email address is tao.zhang@unibw.de.

FABIAN SUHRKE is a Senior Engineer at OSRAM Opto Semiconductors. He is the project leader for fab simulation, real-time dispatching and optimization of production control at the front end in Regensburg. He holds an M.S. degree in mathematics from OTH Regensburg. His email address is fabian.suhrke@osram-os.com.

JONAS MAGER is an Industrial Engineer at Osram Opto Semiconductors in Regensburg. He is focusing on projects to increase fab performance with simulation models and is working on data analyses at the front end production Regensburg. He holds a M.S. degree in industrial engineering and a B.S. degree in electrical engineering from OTH Regensburg. His email address is jonas.mager@osram-os.com.

THOMAS FREY is the Head of the Departments Central Production Control and Industrial Engineering at Osram Opto Semiconductors in Regensburg. He is responsible, among other things, for global reporting, direct materials scheduling and industrial engineering. He received a M.S. degree from Regensburg University and a Ph.D. degree in physics from Paderborn University. Furthermore he holds a MBA from Deggendorf University of Applied Science. His email address is thomas.frey2@osram-os.com.

OLIVER ROSE holds the Chair for Modeling and Simulation at the Department of Computer Science of the Universität der Bundeswehr, Germany. He received a M.S. degree in applied mathematics and a Ph.D. degree in computer science from Würzburg University, Germany. His research focuses on the operational modeling, analysis and material flow control of complex manufacturing facilities, in particular, semiconductor factories. He is a member of INFORMS Simulation Society, ASIM, and GI. His email address is oliver.rose@unibw.de.