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.

TIME BOUND CONTROL IN A STOCHASTIC DYNAMIC WAFER FAB

Tao Zhang
Falk Stefan Pappert
Oliver Rose

Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
Neubiberg, 85577, GERMANY

ABSTRACT

Time bounds are a common constraint in wafer fabs. Releasing wafer into a time bound sequence leads to a tradeoff between capacity loss and yield loss due to violations. Two common approaches to tackle this challenge are static scheduling and dispatching rules. While static scheduling faces problems with the dynamic and stochastic nature of a wafer fab dispatching rules often lack the global perspective causing either unnecessary violations or capacity waste. In this paper, we present an approach taking elements of both these solution approaches to address time bound constraints and compare it to existing approaches.

1 INTRODUCTION

Semiconductor manufacturing is a very complex production environment. Besides other challenging features, time bounds are commonly introduced to the system to safeguard against yield and quality loss due to particle contamination and oxidation. Time bounds basically define a sequence of processing steps, which needs to be completed within a certain time frame. There are different ways to define the start and end points of a time bound. For the purpose of this paper, we will define the beginning of a time bound sequence as the moment a lot/batch leaves the first preparation step. The end is considered to be the moment when the lot/batch enters the last tool needed in the sequence. Figure 1 shows some possible time bound configurations. Time bounds may be set between two consecutive steps, i.e., T1 in Figure 1, or between two non-adjacent steps, i.e., T3. Moreover, sometimes two or more time bounds are required together and may overlap, i.e., T3 and T4. Complex nested time bounds may form in this way. Klemmt and Mönch (2012) presented more details on interactions of time bounds.

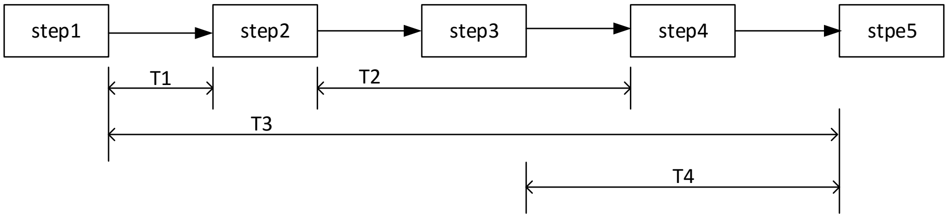


Figure 1: Time bounds and their combinations in wafer fabs.

To ensure successful compliance to these time bounds, tools are reserved or even kept empty. Time bound control strategies therefore try to strike a balance between capacity loss and yield loss due to violations. Basically, time bound control ensures that lots can be finished within time bounds, which is done by deciding whether lots should start into the first step of a time bound sequence. If a lot starts at the first

978-1-5090-4486-3/16/$31.00 ©2016 IEEE

step and will be finished in the given time bound at the last step, the lot may be started; otherwise, the lot should be blocked and wait in front of the first step for sufficient resources, such as operators and tools.

The final purpose of the study is to analyze the performance of the wafer fab under the time bound constraints by using the simulation, so as to improve the operations on shop floor. This paper proposed a time bound control strategy which is then deployed in the simulation. The simulation analysis is presented in our another paper (Pappert et al. 2016). The proposed approach is based on a dynamic planning. A partial plan with a wider global vision and higher adaptability to the dynamic and stochastic environment for the concerned lot is generated before the decision making. The decision will be made according to the partial plan. If, based on this plan, the lot is finished within the time bound, it will be released; otherwise, it will be blocked.

The paper is structured as follows. The literature related to time bound control are reviewed in Section 2. In Section 3, the plan-based approach is presented including batch planning, dispatch planning, and planning with a given batching policy; In Section 4, we use simulation to evaluate the proposed approach and report some essential results. Finally, the paper is concluded in Section 5.

2 LITERATURE REVIEW

Lots of studies are related to time bounds. Some focus on the scheduling problems with time constraints (Yu et al. 2013, Wang, Chien, and Gen 2015, Klemmt and Mönch 2012) while some others study the time bound itself, such as how to select the time bound value and how to predict the probability of reprocessing (Robinson and Giglio 1999). There are also some studies which aim to capacity planning problems with time bounds (Robinson 1998).

However, there are only a few papers related to time bound control problems. Basically two types of approaches are studied so far: decision rules (Scholl and Domaschke 2000) and Kanban-based approaches (Scholl and Domaschke 2000). The decision rules are more suitable to the dynamic and stochastic environment. But the fatal weakness of the decision rules is the lack of the global vision. An extreme case is the basic rule (Reservation), to start lots only if the required tools are available. It will most likely achieve the goal of zero violations, but it will, at the same time, waste capacity. For Kanban-based approaches, the number of lots/batches within the time bounds are restricted. Lots are started only if the current number is less than the number of Kanbans. These approaches result in very smooth lot flows. However, Kanban is problematic when used in areas with strong tool dedication. Kanban as well as other rules will waste significant amounts of tool capacity to achieve zero violations. Kanban can even increase the number of time bound violations in combination with batching if batching is not planned properly before entering the time bound sequence. In some way, scheduling with time constraints can also be considered as one of approaches to solve time bound control problems. The scheduling approach considers time bounds as time constraints and adds them to mathematical programming models which are solved by optimization algorithms with the result of a violation-free schedule. If the schedule can be followed without interruptions, the zero violation goal can be achieved. However, lots arrive dynamically and stochastically, and considering unexpected interruptions, it is impossible to make a whole plan for all lots in advance. Even though a schedule can be calculated, it is often impossible to exactly comply with it in the real world. Similar to our study, Sadeghi et al. (2015) proposed an approach to estimate the probability of satisfying time constraints by a lot waiting for entering time constrained sequences. The decision is made according to the probability. The probabilistic approach is composed of two parts: A disjunctive graph model and a list scheduling algorithm. The probability of satisfying each time constraint is computed by running multiple times the list scheduling algorithm. Nested time bounds are considered in their study, but no batching steps are involved in the time bound sequences.

Considering the discussion above, our study will focus on the time bound control problems with multiple single and batch processing steps, and try to find an approach with a wider global vision and higher adaptability to the dynamic and stochastic environment.

3 PLAN-BASED TIME BOUND CONTROL

3.1 Dynamic Planning

As we have mentioned before, a static schedule runs the risk of not adapting to the dynamic and stochastic environment. In our study, the plan is made for a specific lot at specific moments. Initially, a plan list is built to store all plans which have been made and not finished yet. To decide whether or not to start a lot, we first make a plan for the lot step by step among covered steps according to the plan list and states of the wafer fab at that time. If in the plan the lot is within the time bound, it will be started and the plan will be added into the plan list; otherwise, the lot won't be started and will wait in a queue; Its plan will be discarded. When the lot leaves the time bound area, its plan will be removed from the plan list. Lots which are not constrained by time bounds, are also fitted with a plan. Their time bounds are considered to be infinite.

There are two types of events which trigger plan generation: 1) Lot arrival at an empty blocked queue triggers the creation of a plan for the arriving lot; 2) A tool becoming available with a non-empty blocked queue triggers the creation of plans for lots which may utilize this tool. We call these lots release candidates.

3.2 Dispatch Planning and Batch Planning

There are two different types of plans for the lots: dispatch plan and batch plan. The dispatch plan decides on which tool and in which sequence lots/batches will be processed. The batch plan contains information about which lots or wafers should form a batch and be processed together at a certain step.

For the dispatch planning, there are two sub tasks: lot sequencing and tool allocation. The basic idea behind our approach can be described in the following way. Assuming that at a step, the earliest start time and the processing time of a lot are given or deduced from the previous step. Several tools may be able to process the lot. If we depict assigned jobs on these tools as a Gantt chart, we can find lots of gaps between already scheduled jobs. The dispatch planning for the lot will be the process of finding an earliest suitable gap to put the job into. If the gap exists, the lot will be assigned to the tool which shows the gap. Otherwise, the job will be assigned to the tool which is available first and put to the end. Then, a start time is assigned to the lot. Lots will be processed in the sequence of the start times. The process is shown in Figure 2.

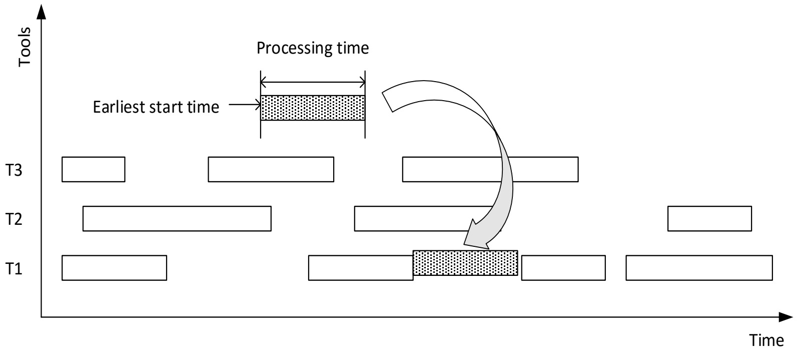


Figure 2: Dispatch planning.

For the batch planning, there is a batch plan list for each batch processing tool group. When we make a batch plan for a lot at a step, we will first try to fill the batches in the batch plan list; after that, if there are still some wafers left, these wafers will form one or more new batches. The new batches will be assigned to one of the tools in the tool group and added in the list. A planned start time and a planned end time are also assigned to each new batch. Once the planned batches are started to be processed, the batches will be removed from the list. Assigned batches to a tool will be processed in the sequence of the planned start times. A batch can only start processing if the previous batches are finished and all of the planned wafers in the batch arrive at the related step.

The procedure is demonstrated in Figure 3. Assuming that there is one lot with 4 wafers and with an earliest start time. The lot can be processed on tools T1, T2, and T3 in batches. B1-B6 are unprocessed batches that were already assigned to the tools. Each batch has a planned start time and a planned end time. When the lot arrives, some batches (B1 and B4) are full while the others are not full. The wafers in the lot will be put into the suitable non-full batches (in this case B5). Due to the batching specifications or start time restrictions, not all non-full batches are suitable, such as B2 and B6. Because only two wafers can be added into B5, the remaining two wafers form a new batch B7 which can accept four more wafers later on. According to the earliest start rule, B7 is assigned to tool T1. The planned start time of B7 is the greater value between the earliest start time and the earlies available time of T1.

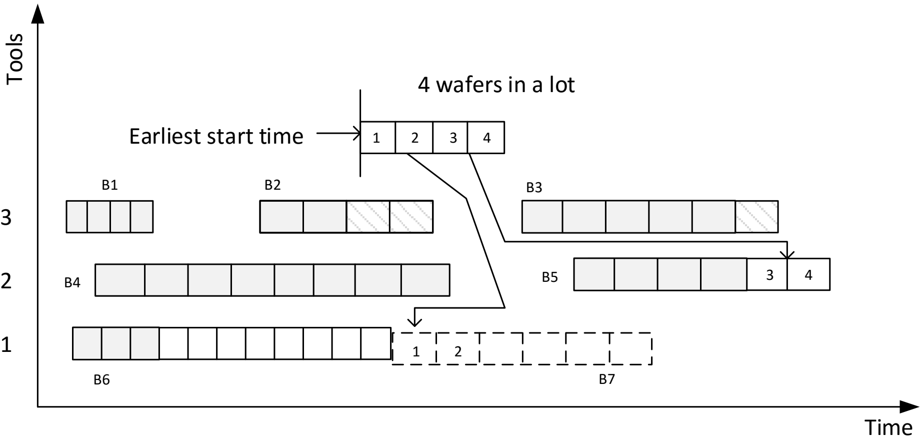
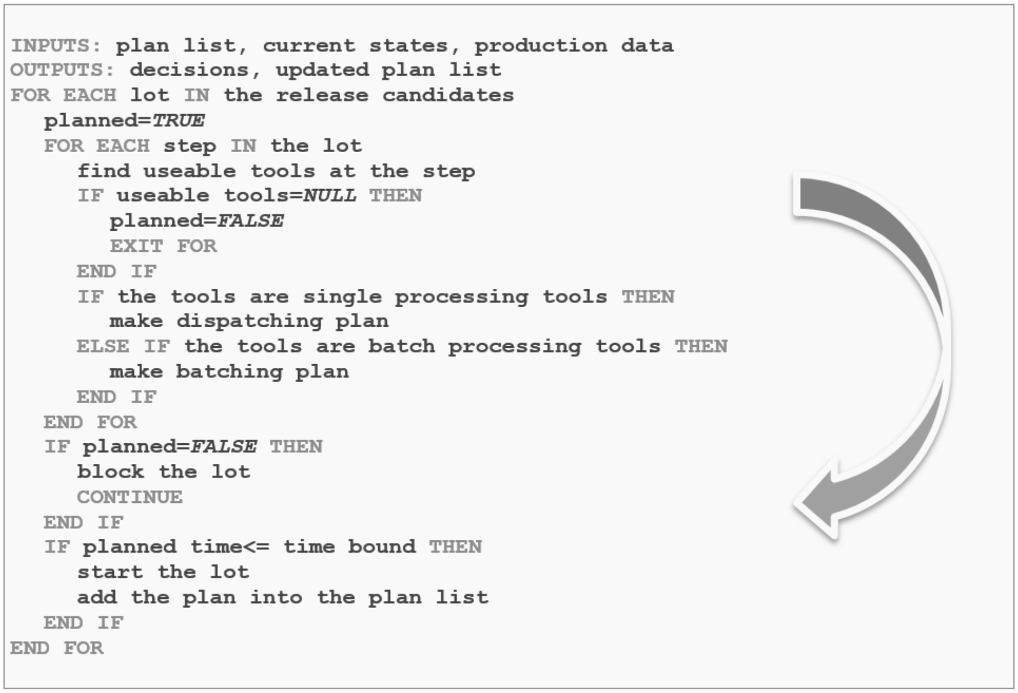


Figure 3: Batch Planning.

3.3 Planning Procedure

The following describes the complete procedure for the dynamic planning. The inputs are the plan list, current state, production data, and the release candidates. The outputs are the decisions whether a lot should start, as well as the updated plan list. In order to deal with stochastic interruptions (such as breakdowns and maintenances), a concept of a “usable tool” is introduced. A usable tool is the tool which is available and idle according to the current plan. If at one step there are no usable tools, the lot will be blocked directly. If a tool becomes unavailable (due to breakdown or maintenance), the tool will be considered as an unusable tool. In order to avoid too many lots are queueing in front of the unavailable tool, a number of maximal planned lots/batches is assigned to each tool. If the number of planned lots/batches is greater than the maximal number, the tool will be considered as an unusable tool. The concept can furthermore prevent lots without time constraints from being started prior to the lots with time bounds.

Once either two events presented in Section 3.1 happens, a plan will be made at each step for each lot. The plan includes a tool allocation plan, a sequencing plan, and a batching plan if the step is a batch processing step. At each step, first we will try to find out the usable tool set. If the set is empty, the planning procedure for this lot will be terminated and move to the next lot. The lot will be blocked. If the set is not empty, a tool will be allocated to the lot at the step and the sequence of the job on this tool is also fixed. A schedule for the lot is made in this way. Based on the schedule, if the lot can be finished within its time bound, the lot will be started; Otherwise the lot will be blocked. Figure 4 shows the pseudocode.



```
INPUTS: plan list, current states, production data
OUTPUTS: decisions, updated plan list
FOR EACH lot IN the release candidates
  planned=TRUE
  FOR EACH step IN the lot
    find useable tools at the step
    IF useable tools=NULL THEN
      planned=FALSE
      EXIT FOR
    END IF
    IF the tools are single processing tools THEN
      make dispatching plan
    ELSE IF the tools are batch processing tools THEN
      make batching plan
    END IF
  END FOR
  IF planned=FALSE THEN
    block the lot
    CONTINUE
  END IF
  IF planned time<= time bound THEN
    start the lot
    add the plan into the plan list
  END IF
END FOR
```

Figure 4: Planning Procedure.

3.4 Planning with Batching Policy

The batch planning that we mentioned above does not follow any batching policy. The batches start processing in the sequencing of the planned start times. Once it is the turn of a batch and all the planned wafers arrived, the batch will start no matter how many wafers are in it. This could waste tool capacity, and increase per wafer cost. In our example, we consider a special case in which the last step of the time bound sequence is a batch processing step and must follow a given batching policy. Due to the bathing policies, like full batching, lots may wait too long to finish batching. As a result, they may violate the time bound.

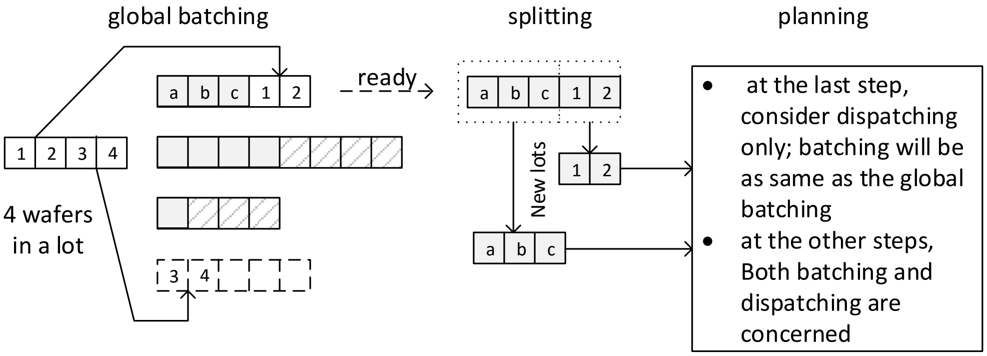

global batching
splitting
planning
a b c 1 2
ready
a b c 1 2
1 2 3 4
4 wafers in a lot
3 4
New lots
1 2
a b c
at the last step, consider dispatching only; batching will be as same as the global batching
at the other steps, Both batching and dispatching are concerned

Figure 5: Planning following batching policy.

Additionally, a global batching and splitting procedure is introduced to deal with this situation. The global batching does not consider the start time, the processing time and the tool allocation. It considers only the batching policies. The batching plan for a lot at the last step is made when the lot arrives at the first step of the time bound sequence. There is a global batching plan list. When a lot arrives, it will check the list to see if its wafers can fill up the batches in the list. If some wafers cannot be added to the started batches, one or more new batches will be created. According to the batching policy, once a batch is ready, the batch will be removed from the list. The finished batches are split into several new lots. Then, we will make plans for these new lots. The last step is only concerned with dispatching. Batching follows the global batching plan. At the other steps, both the batching and dispatching plans are made in the same way as proposed in Section 3.2. The details are described in Figure 5.

4 SIMULATION OF WAFER FAB UNDER TIME BOUND CONTROL

4.1 Simulation under Time Bound Control

In order to validate and evaluate the proposed approach, the time bound controller and a simulation model of the tool sequence are connected. During initialization, the model and the controller are fed with static model data. During the simulation, all state changes are reported to the time bound controller. In the simulation, once a decision is needed, the simulation will pause and request the time bound controller. The controller will in turn generate or query plans and send the results to the simulation. The simulation will adopt the decisions and continue to run. Figure 6 shows the structure of the simulation under time bound control.

On both the controller and the simulation side, an interface is created. The interface on the simulation side facilitates the use of different controllers for simulation. The interface on the time bound controller provides the potential for directly switching to a different simulator or even to the real system. In addition to the plan-based time bound controller, a Kanban-based controller is implemented.

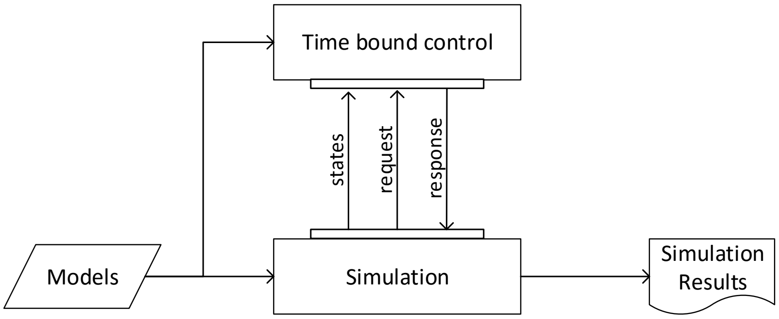


Figure 6: Simulation under time bound control.

The time bound control supports four types of decisions: 1) whether to release a lot into the time bound; 2) which tool a lot/batch should select at a step; 3) Should a lot/batch start on a tool or wait in the front of the tool; 4) in which batches wafers in a lot should be put. All these decisions can be found in the plans. The simulation is actually the process of executing the generated plans.

In addition, it is possible to transplant the proposed approach from the simulator to the real world. However, there are two biggest issues: data collection and compliance with plans. Now all data can be obtained from the simulation. But in real world, the essential data cannot be guaranteed. Moreover, in the simulation, the plan can be completely complied with. Contrarily, the plan may not be complied with in the real world. Both of these two issues will probably influence the performance of the proposed approach.

4.2 Simulation Experiments

The model used for our experiments is based on a generic tool model which we created and parameterized with real fab data provided by one of our industry partners. It represents a few operations leading to a batch coater. For performance and data collection reasons, we only model operations involved in time bounds which end at the batch coater or share equipment with operations that do so. To reduce complexity, the wide range of products is split into product classes. As usual in semiconductor manufacturing, lots may visit batch coating several times during production, with each visit needing a different process. Therefore, a product does not only have a single recipe within the model but usually has several sub recipes to consider. Recipes differ quite a lot, hence we need to consider several tool groups move outs where a time bound sequence may start or which are intermediate tools where lots arriving are already under a time constraint. Apart from normal breakdown and maintenance behavior tools may have different batching options depending on tool and recipe. Time constraints vary from 30 minutes to 48 hours with the short ones being the hardest to fulfill. The main tools, which mark the end of a time bound sequence, are more complex than early and intermediate tools. Main tools in our model are modeled with process dedication and support setup. More details on the used model can be found in Pappert et al. (2016). The number of operators in this experiment are considered to be infinite.

In order to evaluate the proposed approach, several other approaches mentioned in Section 2 are involved in the experiments and compared with our approach. The results are shown in Table 1. Reservation is the rule which starts a lot only if the last step has available tools. Two measurements are considered: violation ratio and average cycle time (CT).

From Table 1 we can see that the proposed approach can reach the zero violation while there are no interruptions. Even though Reservation approach achieves the goal of zero violation, its average CT is longer than the proposed approach. For the model with interruptions, the proposed approach has the second lowest violation ratio. Reservation has the lowest violation ratio, but its cycle time is the longest. If time bound controllers are not applied, the average CT is the shortest, but the time bound violation ratio is the highest. All in all, the proposed approach performs best.

Table 1: Comparisons among different approaches.

Interruptions	Approach	Violation Ratio (%)	Average CT (hour)
No	No controller	23.56	0.38
	Reservation	0.00	0.40
	Kanban	2.91	0.40
	Plan-based	**0.00**	**0.38**
Yes	No controller	25.30	0.46
	Reservation	0.85	1.02
	Kanban-based	5.08	0.52
	Plan-based	**1.31**	**0.49**

5 CONCLUSIONS

Time bound control is an important decision making problem in several industries. Usually, there is a tradeoff between lower violation ratios and longer cycle times. Our proposed plan-based time control makes decisions on the basis of dynamic partial plans which has a wide, more global perspective, as well as higher adaptability to the dynamic and stochastic environment. These properties results in a lower violation ratio while the average cycle time keeps at a lower level. The simulation results show that the approach performs better than other approaches from the literature.

REFERENCES

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*, 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. .

Pappert, F. S., T. Zhang, J. Mager, F. Suhrke, and O. Rose. 2016. "Impact of Time Bound Constraints and Batching on Metallization in an Opto-semiconductor Fab." In *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. .

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.

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. .

Sadeghi, R., S. Dauzere-Peres, C. Yugma, and G. Lepelletier. 2015. "Production Control in Semiconductor Manufacturing with Time Constraints." In *Proceedings of the Advanced Semiconductor Manufacturing Conference (ASMC), 2015 26th Annual SEMI*, edited by unknown, 29-33.

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.

Wang, H. K., C. F. Chien, and M. Gen. 2015. "An Algorithm of Multi-Subpopulation Parameters with Hybrid Estimation of Distribution for Semiconductor Scheduling with Constrained Waiting Time." *IEEE Transactions on Semiconductor Manufacturing* 28 (3):353-366.

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 Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill and M. E. Kuhl, 3630-3641. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

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.

FALK STEFAN PAPPERT is a Research Assistant and PhD student at Universität der Bundeswehr 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.

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.