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.

MODELING OF COMPLEX DECISION MAKING USING FORWARD SIMULATION

Thomas Winkler
Paul Barthel
Ralf Sprenger

GLOBALFOUNDRIES Dresden Module One LLC & Co. KG
Wilschdorfer Landstrasse 101
Dresden, 01109, GERMANY

ABSTRACT

The complexity of simulation models has increased during the last years in semiconductor foundries. Manual and automated decisions have to be modeled in detail to make the right conclusions from them. We describe an approach that uses forward simulation to minimize modeling effort and mimics fab behavior to a high degree. The approach is applied to the problem of controlling time link chains. Results are presented and other applications are discussed.

1 INTRODUCTION

Simulation is important in semiconductor manufacturing due to the complexity of the manufacturing process. More than one thousand process and measurement steps have to be passed by every single wafer. Every product and technology is different and requires other tools and machine setups. Within the last years this complexity has grown to a level which on the one hand requires simulation but on the other hand the development and maintenance effort for the model has increased to a level which is difficult to handle. Especially in foundry business this is caused by a growing number of products, mixed technologies and platforms in one fab. This leads mainly to the following consequences:

- high number of different setups for each customer, product and even version of a product,
- many different systems where restrictions critical for production can be set and
- small-sized wafer structures that lead to additional constraints like time links that increase the complexity to control the material flow.

All of these facts have an impact on the performance of the production line. Naturally, it is difficult to keep track of and understand at which points in the production line improvements or changes are to be made to fulfill customer commitments. Simulation supports identifying and overcoming upcoming problems.

These constraints are difficult to respect in a simulation model and keep the model in sync with the real shop floor. Any change regarding the configuration of the real fab (e.g. dispatch rules) has to be implemented in the simulation model.

Approaches like the APF-FUSION (Applied Materials 2016) engine help to overcome these issues by an automatic integration of dispatch rules into the simulation model. But there are two weaknesses:

- this is only possible if no other technology than the one of this vendor is used at the shop floor and

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

- dispatch rules are optimized to run on a high performance cluster to support near real-time behavior to the shop floor. As a consequence, using simulation is often (depending on the provided computation resource) not necessarily faster than real time.

Accuracy of the simulation model is important to draw the right conclusions from it. The approach that we describe in this paper improves model behavior without time intensive modeling.

In Section 2, we will discuss the problem in detail. Section 3 describes the new approach. In the following, we show how the approach can be applied by implementing it to control time links. Section 5 will summarize and give some ideas of possible future applications of this approach. Because this paper covers several aspects, related literature is discussed separately in each section.

2 PROBLEM DISCUSSION

Simulation model data in semiconductor industry consists of different types of data (Table 1). Decisions that are made during simulation based on implemented rules are usually made with data of types 1 and 4. For example, a machine becoming available for processing new material triggers a dispatch decision that considers flow information about which material can be processed (Type 1) and selects the most important material according to rankings that depends on the current situation (Type 4). During a dispatch decision, predictions about downstream operations might be used to improve the decision making. A prediction about the future is made and the consequence of a decision that needs to be made at that time is taken into consideration. We will come up with an example in Section 4 to clarify this in detail. Type 5 data is part of any simulation run with random events. We will describe its usage in the next section.

Table 1: Types of simulation model data.

Type	Description	Example
1	Master data that is not changed during simulation runtime.	Flow related information.
2	Data that is valid at simulation start time. After initialization of simulation, this data is not required anymore.	Current position of material in the fab or the current state of machines.
3	Change lists about model data of type 1 that is manipulated during specific points in simulation time.	Additional machines that come into production later on.
4	Highly dynamic data that is only generated during the simulation.	Sorting of dispatch lists or other decisions that the simulation model has done for a later time in the simulation. Current state of tools, upcoming material in a currently processed cascade.
5	Random number generators used to model stochastic events.	One or more random number generators are initialized with random seeds at simulation start to model machine failures.

In foundry business, the ability to take future consequences into consideration is needed:

- by complex dispatch rules or
- by an operations engineer who needs to make decisions on the shop floor or by the Line Control Department trying to optimize the material flow to maximize performance or get customer orders fulfilled in time.

Implementation of complex dispatch rules within the simulation is possible (as described in Section 1) but requires high effort. For modeling manual decisions, specific business logic has to be implemented in the simulation model that tries to model and predict them. From our experience, this approach does not pay off and is highly fragile and error-prone.

One of the characteristics of simulation models is that a wrong prediction at one point in the whole simulation run can lead to very bad overall results. For example, material getting stuck at one point in the simulated production line may render all downstream predictions wrong and any use of them might lead to wrong decisions in reality. This could apply, for example, to a set of machines that is prepared for an expected upcoming material quantity. In reality this quantity will be much higher than predicted by the simulation.

To avoid complex modeling, simple simulation models can be used for analysis of complex fab behavior as described in Sprenger and Rose (2010). In contrast, selecting only some of the critical parts of the fab for detailed modelling of decision making while ignoring other “less critical” parts may neglect important correlations between the individual complexities. The approach discussed here adds to prediction quality without adding to modelling complexity.

3 FORWARD SIMULATION APPROACH

We will start by describing an approach to overcome the aforementioned weaknesses (Figure 1) before it is used to control time link chains in a semiconductor foundry in Section 4. The basic concept is a simulation within the simulation whose results are fed back and used for decision-making:

1. Let $t \in \mathbb{R}$ be the current simulation time and $fi > 0,\ fi \in \mathbb{N}$ is given. The main simulation will be interrupted whenever $t \in FI$, where $FI = \{l * fi\}, l \in \mathbb{N}$.
2. At that point, a deep copy of the complete simulation model, including data of types 1, 3, 4 and optionally 5, is created.
3. The copied model continues simulating the behavior of the fab within time $fh \in \mathbb{N}, fh \geq fi$ (e.g. for the next few hours or days).
4. A defined set of results is fed back from the copied simulation to the main simulation and the copy is removed from memory.
5. The main simulation continues and decisions are made using the input of the copied model run.

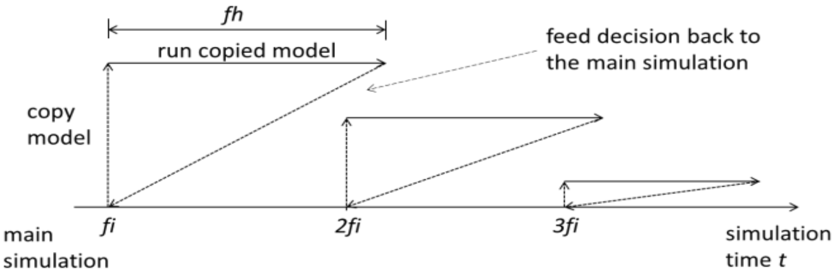

Figure 1: Main approach.

3.1 Advantages

In Section 4 a complete example of how this can be used for easily modeling complex behaviors in a shop floor is given. However, we want to give some remarks, before.

Complex decision making is often done by calculating future machine loading (e.g. Scholl and Domaschke 2000) and generating schedules for material. This information is used to predict future behavior and to make a decision at this time. However, that requires complex internal modeling of some parts of the supply chain and includes many assumptions. If many products and technologies are involved, there are many interactions with other parts of the line. This is often modeled by constant incoming streams of material without having a good prediction. The approach discussed here overcomes these issues in the following ways:

- A copy of the whole simulation model has full details. No assumptions have to be made.
- Besides configuration parameters that will be described in Subsection 4.2, the only thing that has to be modeled is which data is used to make a decision and how it is executed afterwards.

3.2 Applications

We want to give some application examples before a concrete application is discussed in detail in Section 4. The copied simulation model can use the same random seed (Type 5 data) as the main simulation. This leads to the situation that, beside of the actions that are taken from the copied model, the main simulation behaves exactly like its copy regarding machine down times or other stochastic influences. Alternatively, the copy can be started with new seeds to evaluate the behavior under uncertainty. Importance of evaluation under uncertainty can be found in Sprenger and Mönch (2012). For this, the described approach also enables to run many replications with different random seeds to improve the decision making (Figure 2).

In real production systems, any control system accounts buffers for uncertainty. If the simulation were not doing so, as well, it would assume that the fab could be controlled perfectly. As a result the behavior of the model would be much better than in reality.

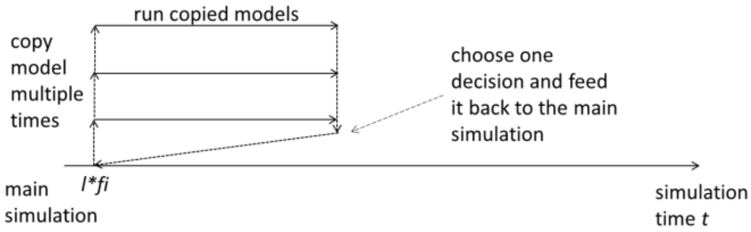
run copied models
copy model multiple times
choose one decision and feed it back to the main simulation
main simulation
l*fi
simulation time t

Figure 2: Multiple runs and final decision making.

The forward simulation approach can also be used for continuous improvement by iteratively changing the control mechanism and re-evaluating with a new copy (Figure 3). This can be used for a step-by-step improvement of a decision.

Another use case of this approach is to make decisions and control the production system directly. The simulation model is often a well-maintained model of the reality. Thus, if a decision is made within the simulation and it is proven that it can handle a specific degree of uncertainty, it can be used to control the fab by feeding the decision back to the production system for execution. In Hongtao and Zhangb (2012) an approach for stepwise decision making for controlling a semiconductor fab is described.

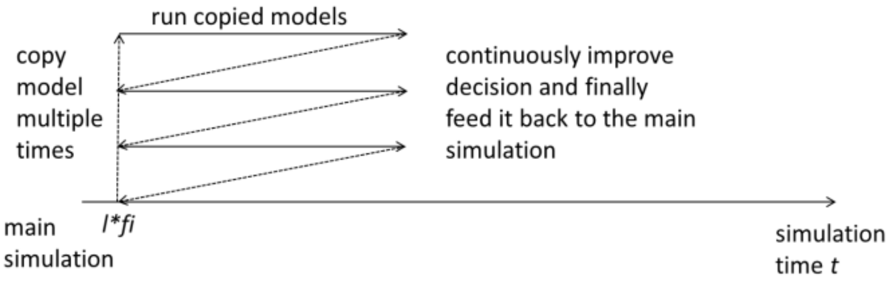
run copied models
copy model multiple times
continuously improve decision and finally feed it back to the main simulation
main simulation
l*fi
simulation time t

Figure 3: Continuous improvement of decisions.

4 AUTOMATIC CONTROL OF TIMELINKS

In semiconductor manufacturing, material is being processed according to a process flow $\mathrm{p_o} \ldots \mathrm{p_k}$, $\mathrm{k} \in \mathbb{N}$. Time links define a maximum time $\mathrm{tl} > 0,\ \mathrm{tl} \in \mathbb{N}$ between process end at operation $\mathrm{p_m}$ and process start at $\mathrm{p_n}$ where $m < n$ must apply. The time within the time link is defined by the process time pt and waiting time wt at all operations within the time link and the waiting time at the operation $\mathrm{p_n}$ so that if

$$\mathrm{wt}(\mathrm{p_n}) + \sum_{\mathrm{x=m+1,x\in \mathbb{N}}}^{\mathrm{n-1}} \mathrm{wt}(\mathrm{p_x}) + \mathrm{pt}(\mathrm{p_x}) \leq \mathrm{tl}$$

is valid, the time link has been passed without violation.

Time links are often connected to time link chains (Figure 4). That means at operation $\mathrm{p_n}$ a new time link can start. Detailed description of different time link variants can be found in Klemmt and Mönch (2012). Our approach is independent from which variant applies. Therefore we will not discuss all types here.

There are two ways to prevent material from violating the constraints within a chain:

- time link material can be handled with higher priority against non-time link material (e.g. Scholl and Domaschke 2000) and
- material can be stopped at the first operation (time link gate) of a chain.

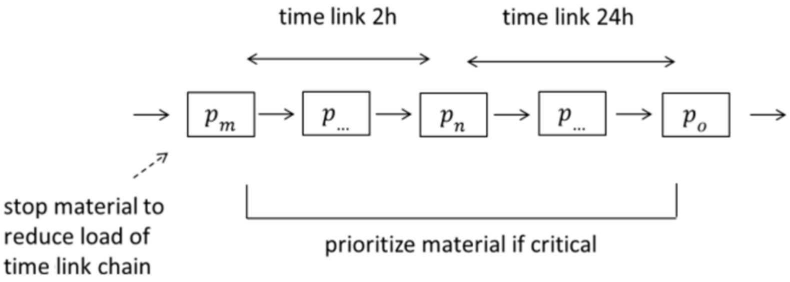
time link 2h
time link 24h
p_m
$p_{\ldots}$
p_n
$p_{\ldots}$
p_o
stop material to reduce load of time link chain
prioritize material if critical

Figure 4: Example of a time link chain.

In the first approach a criticality factor $0 \leq \mathrm{cf} \leq 1, \mathrm{cf} \in \mathbb{R}$ defines which material is critical and prioritizes it if this ratio of the time link constraint is reached. This leads to the fact that material for which

$$\left(t - t(p_{m+1})\right) \geq cf * tl$$

is valid, is preferred against all non-critical material. t is the current time within the simulation and $\mathrm{t(p_i)}$ the time when material enters operation $\mathrm{p_i}$.

Prioritization is done compared to non-critical time link material but also based on the remaining time until a violation will occur. This means:

- non-critical time link material is handled similar to non- time link material,
- critical time link material is processed before all other material and
- critical time link material among one another is prioritized depending on the remaining time when a violation will occur: $tl - \left(t - t(p_{m+1})\right)$.

We use this approach to prevent violation. A similar method is used for example in Scholl and Domaschke (2000).

For small technology nodes, the restrictions are much tighter (typically around 1 to 72 hours). Depending on the variability and other tool characteristics, the only way to prevent violations is to stop material before being processed at the first operation of the time link chain. Therefore rules are developed to make the decision which material will be fed into the time link area and which material has to wait. One suggestion is made by Levy, Burda and Stahlecker (2010).

This is a typical example of the complexity of dispatch rules. Implementing this rule within the simulation is possible but requires much effort.

4.1 Decision Approach

In the copied simulation model run, the following information about material within time links is tracked:

- material has entered time link chain
- material has left time link chain
- material has violated one of the constraints within the time link chain

This information is fed back to the main simulation, the copy is removed from memory. For stopping material at the time link gates operations in the main simulation we implemented two approaches:

- Blacklist approach: put all material that has violated one of the time link constraints during the copied simulation run to a blacklist and prevent this material to be released into the corresponding time link area. Material not on this list is released into the chain.
- Whitelist approach: put all material that successfully passed the time link without violating a constraint to a white list. Only material on this list is released into the chain.

4.2 Configuration Parameters

We evaluated which configuration parameters lead to good results and derived the following parameters where the approach mimics the behavior of a real fab:

- forecast horizon $fh = 240h$
- forecast interval $fi = 9h$
- time link criticality factor $cf = 0.75$

The fact that we set fh $\gg$ fi is caused by long time link chains that have to be evaluated to be passed successfully or not. We make decisions only for the next few hours (fi) but if a decision is made in the primary simulation to allow material to be fed into the time link chain, this might lead to a violation of a time link constraint a few days later.

However, the simulation is slowed down by factor $\frac{\text{fh}}{\text{fi}}$. Some time for copying a model comes on top so that the simulation is slowed down approximately by factor 30 using these parameters. Even so, the simulation is still much faster than real time and we do not see much negative impact of this fact in practical use. If real dispatch rules are used like in the APF FUSION approach, the simulation is slowed down close to real time level which is not acceptable for our purposes.

4.3 Blacklist vs. Whitelist Approach

We will take uncertainty into account by copying type 5 data for all following results. How well time link chains are controlled is mainly determined by

- the time it takes for material to pass a chain,
- the maximum throughput of the chain and
- number of time link violations that occur.

Figure 5 shows a coarse comparison of the different approaches. Optimal values are located at the outer corners. The Whitelist approach reduces time link violations close to zero whereas the Blacklist approach leads to a few violations but much less than without a gate control using only time link prioritization strategy. However, this approach leads only to slight degradation of the capacity and speed of the chain and mimics the real behavior of a fab. Therefore we use the Blacklist approach in our regular simulation models and for the following evaluations.

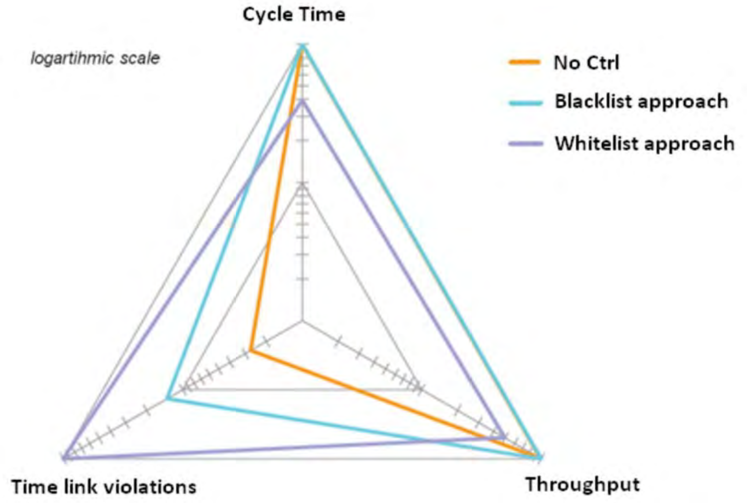

Figure 5: Comparison of the different approaches.

4.4 Improvement of the Blacklist Approach

To benefit from the higher throughput and lower cycle time offered by the blacklist approach while not causing too many time link violations, we implemented a simple approach to modify the decision that is used for controlling which material is put into the time link chain.

We add material to the blacklist that passed the time link chain but was close to violating the constraint. Normally, all material that has passed successfully is excluded from the list. The modification adds material to the blacklist that only has a specific relative remaining time after passing the time link successfully to reduce the risk that it may still violate due to uncertainty.

Figure 6 compares for a full fab model the amounts of material that successfully passed a time link, violated a time link constraint and became critical but still passed successfully. A scenario without active time link control is shown and compared to the approaches with active control: one without the modification described (0%) and three with different remaining time required after passing a time link for material not to be blacklisted.

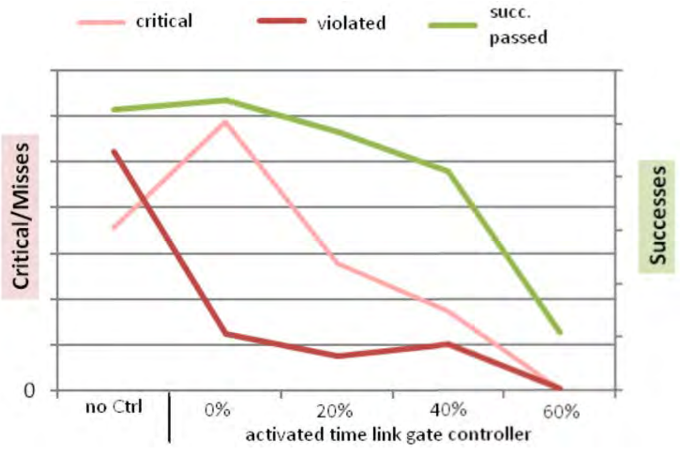

Figure 6: Behavior of full fab model.

This clearly shows that our gate control approach works. The number of time link violations is reduced significantly at similar throughput. The slight increase is caused by changes of the material flow within the model due to the activated logic. We used the real fab model with a few thousand machines and hundreds of products with around one thousand process steps each for this evaluation instead of a theoretical model to assess the approach under realistic conditions. Simulation horizon is one quarter and we did 10 replications. However, the conclusion that this logic increases throughput of the whole fab would be incorrect.

The amount of material that gets critical but passes without violation increases significantly. That indicates that the control mechanism works well and puts more material into the time links while the number of violations is reduced to a high degree.

For further reduction of time link violations, a high amount of material has to be blacklisted resulting in increased throughput impact. Until 40% reduction of allowed material the violations persist at the same level. At 60% the violations are close to 0 but throughput is severely reduced.

Figure 7 shows the impact at one single time link chain within the reoccurring metal layers in the production flow. The situation looks different because

$$\mathrm{tl} - \sum_{x=m+1, x \in \mathbb{N}}^{n-1} \mathrm{pt}(\mathrm{p}_x)$$

is very low, there is no big margin for waiting time and this area has a high variability where uncertainty has a huge impact. The time link control can only slightly reduce the number of violations. The violations can be reduced only with significant reduction of material that is fed into this area.

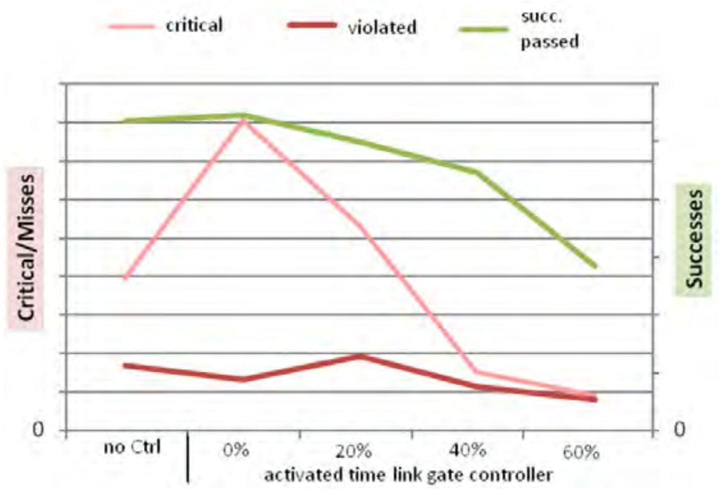

Figure 7: Behavior of one critical time link chain.

4.5 Usage in Predicting Future Fab Behavior

Machines that limit fab performance often occur within time link chains. Without a working control mechanism that stops material at the first operation of the chain, all material at this operation is instantly moved to this bottleneck machine group within the time link. The described approach is regularly used in predicting fab behavior at GLOBALFOUNDRIES and it improves the upcoming material forecast.

In addition, we can compare the predicted time link violations within the simulation with reality. This gives us a good idea about how good the control approach implemented in the dispatch rules are and at which points in the line a variability reduction can help. Reducing uncertainty in important time link chains can reduce the number of violations.

5 OUTLOOK

It has been shown that the described approach can be applied to complex problems and improves forecast accuracy. Furthermore, it can give additional insights into the problems of the real fab and at which points an improvement makes sense.

Besides that, we think about other problem areas where this approach can be very useful. One application where we plan to use it is to use forward simulation to improve machine maintenance planning. For that, we make multiple copies of the model and vary the planning in each of them. After that, we choose one and feed it back to the main simulation to assess them under uncertainty.

REFERENCES

Applied Materials. *APF FUSION – A Dispatching and Capacity Analysis Solution that uses APF RTD Rules Directly within an AutoSched AP Simulation Model.* http://www.appliedmaterials.com/global-services/automation-software/apf-fusion [Accessed March 6, 2016].

Hongtao, H., and H. Zhangb. 2012. “A Simulation-based Two-stage Scheduling Methodology for Controlling Semiconductor Wafer Fabs.” In: *Expert Systems with Applications* 39(14):11677–11684.

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, 2173-2182. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Levy, J., R. Burda, and T. Stahlecker. 2010. “Method for Determining Amount of Product Released into a Time Sensitive Operation.” In *Proceedings of the 2010 Winter Simulation Conference*, edited by B.

Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 2523-2530. 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.

Sprenger, R., and O. Rose. 2010. “On the Simplification of Semiconductor Wafer Factory Simulation Models.” *Conceptual Modeling for Discrete-Event Simulation*. Boca Raton, FL, USA: Taylor&Francis, 451-470.

Sprenger, R., and L. Mönch. 2012. “A Methodology to Solve Large-Scale Cooperative Transportation Planning Problems.” *European Journal of Operational Research* 23(3):626–636.

AUTHOR BIOGRAPHIES

Thomas Winkler is an engineer in the line analysis department of GLOBALFOUNDRIES Fab 1 in Dresden. He received a master’s degree in Applied Information Technologies at Dresden University of Applied Sciences. His research interests include capacity planning, heuristics for multiple-constraint problems, data visualization and simulation. His email address is thomas.winkler@globalfoundries.com.

Paul Barthel is an engineer in the line analysis department of GLOBALFOUNDRIES Fab 1 in Dresden. He received master’s degrees at Magdeburg University and Rose-Hulman Institute of Technology. His research interests include simulation, production control, automated data analysis and optimization. His email address is paul.barthel@globalfoundries.com.

Ralf Sprenger is manager of the line analysis department of GLOBALFOUNDRIES Fab 1 in Dresden. He received a Ph.D. from the Department of Mathematics and Computer Science at the University of Hagen, and a master’s degree in computer science at Dresden University of Technology. His research interests include industrial engineering in semiconductor manufacturing and optimization. His email address is ralf.sprenger@globalfoundries.com.