Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

BUILDING A DIGITAL TWIN OF THE PHOTOLITHOGRAPHY AREA OF A REAL-WORLD WAFER FAB TO VALIDATE IMPROVED PRODUCTION CONTROL

Patrick C. Deenen

Industrial Technology and Engineering Centre
Nexperia
Jonkerbosplein 52
Nijmegen, 6534AB, THE NETHERLANDS

Rick A.M. Adriaensen

Department of Industrial Engineering
Eindhoven University of Technology
PO Box 513
Eindhoven, 5600 MB, THE NETHERLANDS

John W. Fowler

Department of Supply Chain Management
Arizona State University
P.O. Box 874706
Tempe, AZ 85287-4706 USA

ABSTRACT

Since the photolithography area is generally the bottleneck of a wafer fab, effective scheduling in this area can increase the performance of the complete fab significantly. However, the potential benefit of proposed solution methods is often validated in a static and deterministic scheduling setting, while the manufacturing environment is dynamic and stochastic. In this paper, we build a discrete-event simulation model based on real-world data of the photolithography area, which can be used to accurately determine the performance of new scheduling solutions. A case study at a global semiconductor manufacturer is presented. The simulation model, a so-called digital twin, captures the vast majority of the stochastic behavior such as the arrival of jobs, processing times, setup times and machine downs. In addition, a dispatching heuristic is developed to replicate the current practice of production control. Both the simulation model and the dispatching heuristic are validated and shown to be accurate.

1 INTRODUCTION

The photolithography area contains the most capital-intensive machines of a semiconductor wafer fabrication facility (wafer fab). It is crucial to ensure a high utilization of these machines to minimize the number of machines needed. Therefore, the photolithography area is generally the bottleneck of a wafer fab. Effective scheduling in this area can increase the performance of a fab significantly. A lot of research has been done in this field, resulting in a great variety of solution methods for this problem. However, it is difficult for a semiconductor manufacturer to find the solution which fits their manufacturing facility best, because each wafer fab has different scheduling constraints or objectives. Besides, many solutions methods are based on a deterministic and static scheduling problem which is extracted from the stochastic and dynamic real-world. The potential benefit is quantified in the deterministic and static setting as well and does not accurately reflect the real-world benefit. A realized production schedule from the real-world is shown in Figure 1 and illustrates the stochastic behavior of a photolithography area. Semiconductor manufacturers are very reluctant to test new production control methods in a running manufacturing environment, especially in photolithography, because it can severely impact the performance of the complete wafer fab. Without an accurate estimation of the potential benefit, it is hard for semiconductor manufacturers to justify the business case of implementing a new production control solution. Thus, there is a need for an offline validation method which can accurately quantify the potential benefit of new solution methods.

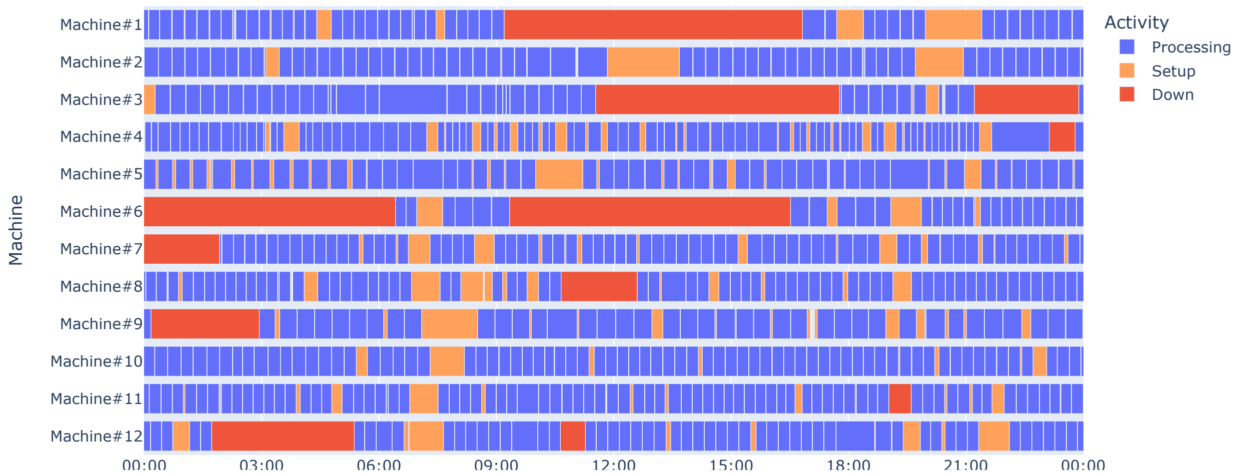


Figure 1: Visualization of the real-world production on machines in a photolithography area.

In this work, we propose a discrete-event simulation (DES) model which mimics the photolithography area of a wafer fab. A real-world case study at one of the wafer fabs of Nexperia, a global semiconductor manufacturer, is presented. The simulation model is referred to as a digital twin, because it is based on real-world production data and can be synchronized with the current real-world state. The digital twin is comprised of several elements: the machines, the dispatcher and the lot generator. The machines capture the behavior of the real-world photolithography machines, so-called steppers, such as stochastic machine-dependent processing times, stochastic machine-and-sequence-dependent setup times and random machine downs with stochastic length. The dispatcher dictates which jobs are dispatched from the queue onto an idle machine. We develop a heuristic which captures the current practice of the production control at Nexperia, but this element can be replaced with other production control methods to validate the potential improvement of each of these methods. Finally, the lot generator mimics the arrival of batches of maximum 25 wafers, i.e. lots, from the real-world measured arrivals at the photolithography area.

The importance of scheduling in a semiconductor facility has been known for many years and there exists a rich literature on this topic. A detailed overview is given by Mönch et al. (2011) and another overview of only solution methods using dispatching rules is given by Sarin et al. (2011). Although many dispatching or scheduling solutions are developed for the photolithography area, none of them are validated in a dynamic and stochastic environment which realistically mimic the real-world. Instead, it is often chosen to validate it in a deterministic and (sometimes) static environment or directly compare it with the realized production schedule from the real-world. Geiger et al. (2006) and Cakici and Mason (2007) both consider the problem with deterministic processing times and without setup times. The experimental setting is based on comparing the static schedules of the novel solution methods with benchmark dispatching rules or heuristics, respectively. Similarly, Ham and Cho (2015) compare Gupta and Sivakumar (2006) consider the problem of allocating jobs to single machines. Although they do consider stochastic processing times, setup times are not considered and the approach is not analyzed for the parallel machine case. Doleschal et al. (2013) solves the lithography scheduling problem with a three-stage integer linear programming (ILP) method and this approach is validated on test instances. These test instances do mimic the dynamic behavior, but are generated randomly. Besides, the solution does not incorporate many of the real-world constraints such as reticle duplicates. Ham (2018) and Bitar et al. (2016) both consider deterministic processing and setup times and validate their approaches with randomly generated test instances. Ham and

Cho (2015) does include processing and setup times obtained from the real-world, but are still deterministic. Finally, Janssen (2019) introduces a practical and computationally efficient ILP to solve the problem and compares the solution to a static schedule created from the real-world schedule.

The main contribution of this work to literature can be summarized as follows: we provide a digital twin of the photolithography area to accurately validate the performance of production control methods. Contrary to the static and deterministic setting often used for validation in other literature, we propose a dynamic and stochastic environment which is able to realistically quantify the real-world potential benefit of new production control methods. Although the digital twin in this work is tailored for the situation at Nexperia, the potential application extends to many other wafer fabs by using the production data and dispatching logic of the corresponding wafer fab.

The remainder of this paper is structured as follows: we will start by explaining the simulation model, the used data for that model and the dispatching heuristic to mimic current practice in Section 2. After that, we will discuss the results in Section 3 to verify the accuracy of the simulation model. Finally, the conclusions and recommendations for future work are given in Section 4.

2 SIMULATION MODEL

The discrete-event simulation model is modelled with the C# Simulation Library (CSSL) (Adan and Deenen 2021). Different experiments can be executed by the simulation model. One experiment consists of a predefined number of replications using one specific setting of the simulation model. There are several settings which can be set by the user, such as the production control method, the number of replications, the start date and the end date. In this work, we focus on merely one production control method: current practice. This will be the benchmark for future work, where other production control methods can be implemented in the simulation model. Since the simulation contains stochastic behavior, the number of replications can be set. Data is collected over a time period of four consecutive months, so each period within these four months can be simulated by setting the start and end date.

2.1 Architecture

The architecture of the simulation model is illustrated in Figure 2. The classes *LithographyArea*, *LotGenerator*, *Dispatcher* and *Machine* are created, each with their own functionalities. Here the *LithographyArea* class connects all the individual classes and records all the output data for each replication.

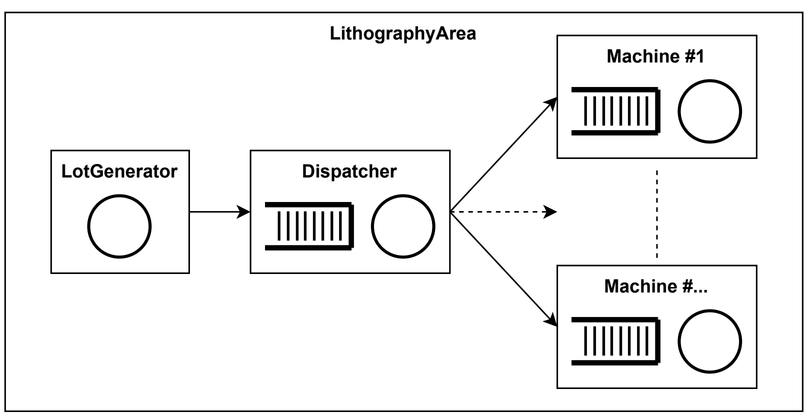


Figure 2: Schematic overview of the architecture of the simulation model.

The *LotGenerator* class generates each lot used during a replication and sends each lot to the *Dispatcher* class at its arrival time. Many different product types are manufactured simultaneously. Each product type consists of a predefined routing, consisting of production stages which have to be performed at dedicated work areas. Wafers are build layer by layer and each layer needs a photolithography stage. Thus, one product type has multiple production stages which visit the photolithography area. We define a lot type as a unique combination of product type and production stage, since this combination determines the characteristics of the processing in photolithography, such as processing speeds, setup times and machine eligibilities. Each lot arrival is characterized by a lot type and operational due date, corresponding to the measured lot arrivals in the real-world.

The *Dispatcher* class contains the queue in which all lots are stored when they have arrived in the system and are waiting to be processed by a machine. When a machine becomes available, the dispatcher determines which lot is sent to that machine. This decision depends on the solution method used in the experiment. When a lot is send to a machine, the usage of the required auxiliary resources is updated and if the maximum capacity of an auxiliary resource is reached, no lot which need that auxiliary resource can be dispatched until it becomes available again. The dispatcher also accounts for the machine eligibilities and will not dispatch a lot to a machine which is not eligible to process it.

An instance of the *Machine* class is created for each stepper in the photolithography area and replicates the processing of each lot. When a machine becomes available, it receives the next lot from the queue of the dispatcher. It then samples the setup time from the corresponding distribution depending on the previous lot type, the dispatched lot type and the machine. When the setup time has passed, the machine samples the processing time of the dispatched lot from the corresponding distribution depending on the lot type and the machine. When the processing time has passed, the lot is finished and the next lot can be dispatched. After finishing a lot, the used auxiliary resources is released. If no lot is send by the dispatcher, due to, e.g., an empty queue, the machine waits until the dispatcher sends a new lot. Each machine m can go down after finishing a lot with a probability $P(down_m)$. If a machine goes down, a machine-dependent downtime is sampled from the corresponding distribution. After this downtime, the machine requests a new lot from the dispatcher.

2.2 Input Data

To build the data-driven simulation model, data is extracted from the manufacturing execution system (MES) for four consecutive months of production. This data includes the recipe enablements, lot arrivals, processing times, setup times and machine downs. Each of them will be explained in more detail in this section.

2.2.1 Recipe Enablements

Each lot type requires a specific recipe to be processed on a stepper. The recipe determines the machine settings such as stepping speed, light intensity and total number of flashes per wafer as well as what reticle is needed. A stepper has to be qualified to process a certain recipe. Due to quality reasons or machine capabilities, only a subset of all machines are capable to process each recipe. This dictates the machine eligibilities in the production control. Thus, a recipe can be enabled or disabled on a certain machine which is referred to as *recipe enablements*. These recipe enablements change over time in the real world and are therefore logged every week.

2.2.2 Lot Arrivals

The start and end of each processing step on a lot is logged in the MES and is referred to as the track-in and track-out respectively. The lot arrivals at photolithography are retrieved by filtering on the track-out of the step prior to the photolithography step. When simulating a certain time window, lots that arrived before the end of the time window and departed after the start of the time window are generated at their arrival time. All other lots are not included in the experiment. By combining several data sources, one input file for the simulation model is generated which contains following information for each lot arrival:

- Arrival time
- Departure time
- Number of wafers
- Operational due date
- Layer type
- Required reticle
- Required recipe
- Priority

The operational due date is the due date set for the single photolithography step. Each wafer has a single final due date on which the wafer is supposed to be finished. To keep wafers on track, a operational due date is set for each single step in the production process. The layer type indicates on which layer the current wafer is. As mentioned before, wafers are produced layer by layer. Although there exist many different product types, many of them contain similar layer types. These layer types determine certain processing characteristics such as the setup times (see Section 2.2.4) and are used in the current practice of production control (see Section 2.3). The priority can be either normal or high. Lots with high priority are so-called *hot lots* and have to be given priority over the normal lots.

2.2.3 Processing Times

The realized processing times of all lots in the real world are logged in the MES and are used to generate the stochastic processing times in the simulation model. The logged processing times are grouped per machine-reticle combination. This grouping is chosen, because (1) machines have different speeds and (2) analysis showed that the processing times on a single machine are mainly determined by the used reticle. The processing times are removed from the data set if: lots contained less than 25 wafers or the machine-reticle combination contained less than three samples. This results in a set of processing times for each machine-reticle combination (with three or more recorded samples of 25 wafers). Each set is translated in an empirical distribution, where each sample has equal probability. The distribution is positive skewed due to machine hiccups, which do not occur frequently but increase the processing times significantly. An example of the processing time distribution of a unique machine-reticle combination is shown in Figure 3.

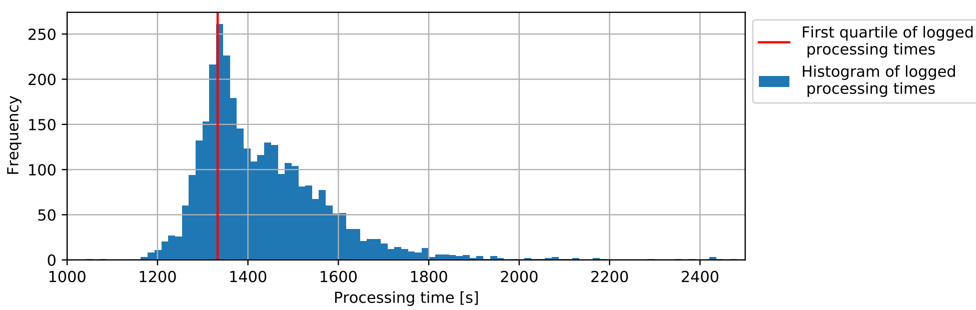


Figure 3: Processing time distribution of a unique machine-reticle combination.

In practice, not all eligible machine-reticle combinations were used in the real-world in the considered time period for data collection. Therefore, processing times of a several machine-reticle combinations were unknown. To ensure a proper functionality of the simulation model, the machine-reticle combinations with unknown processing times are made ineligible. This assumption is valid, because we merely want to mimic current practice of production control. If one wants to consider other production control methods, it is advised to explore possibilities to retrieve these missing processing times. A possible direction would be, to use machine learning to predict a generic distribution for the missing machine-reticle combinations.

When a lot contains less than 25 wafers, the sampled processing time p_{jm}^{25} is scaled linearly by the number of wafers in the lot according to equation (1).

$$p_{jm} = \frac{n_j}{25} \cdot p_{jm}^{25} \tag{1}$$

where, p_{jm} is the processing time of lot j on machine m, p_{jm}^{25} is the sampled processing time of lot j on machine m if the lots contains 25 wafers and n_j is the number of wafers in lot j. This assumption is validated with machine data, which contains the processing times of individual wafers. Wafers are processed sequentially and it is observed that processing times per wafer are fairly constant.

2.2.4 Setup Times

The non-productive times are the times between the processing of two subsequent lots. These non-productive times can be categorized as either a setup time or a down time and these are both logged in the MES. Based on analysis and expert knowledge we identify three groups of setup types: setup time between lots of the same layer type using the same reticle, setup time between lots of the same layer type using a different reticle and setup time between lots of different layer types using a different reticle (a unique reticle is never used among different layer types).

The realized setup times of the real world in the four considered months are logged. These setup times are grouped by machine and setup type. A machine setup requires manual intervention of a shop-floor operator. A single operator has to serve multiple machines and setups are often delayed because the operator was not in time. This causes the setup times to be very stochastic but also time-dependent. Operator configuration may change over time to prioritize throughput on specific steppers. Therefore the setup times are also grouped per month. This results in a set of setup times for every machine, each setup type and each month. Each set is translated in an empirical distribution where each sample has an equal probability.

2.2.5 Machine Downs

The second category of non-productive times between lots are the machine downs. When a machine goes down, a maintenance activity takes place and these activities are logged. All non-productive times in which one or more maintenance activities have taken place are categorized as down times. These down times are grouped per machine and month. The monthly grouping is to account for the fact that the frequency of these machine downs are relatively low, but their impact on the performance is high. Since we want to mimic the real-world fab, a monthly grouping ensures less stochastic deviations from the real-world in terms of machine downs. Each set of down times is used in an empirical distribution, where each sample has equal probability. Whether a down time takes place depends on the probability of a machine m going down in month x; ($P(down_{mx})$). This probability is calculated according to:

$$P(down_{mx}) = \frac{n_{mx}^{down}}{n_{mx}^{total}} \tag{2}$$

where, n_{mx}^{down} is the number of non-productive times identified as down times on machine m in month x and n_{mx}^{total} is total number of non-productive times on machine m in month x.

2.3 Current Production Control

To accurately simulate the real-world production, we also need to mimic the current practice of production control. This logic will be captured in the *Dispatcher* class (see Section 2.1) and can be the benchmark for other production control methods. In consultation with experienced shop-floor operators we identified the current work flow of production control. The current work flow is a combination of some manual decisions (from both managers and operators) and certain priority rules. Since different managers or operators can make different decisions, the production control is not completely consistent. This stochastic component of the manual decisions is not included, because (1) it is hard to capture and (2) the stochastic influence is limited. Instead, these manual decisions are analyzed and translated into deterministic rules. This results in a dispatch-based heuristic, which will be explained next. Due to confidentiality reasons, we cannot describe this heuristic in detail, but a general description will be given.

A graphical overview of current production control is shown in Figure 4. Every 24 hours, it is determined what number of wafers should be processed in the next 24 hours of each specific layer type, so-called production targets. These targets depend on the work-in-progress (WIP)-balance, which are the number of lots present at the other production steps in the fab. Products have to go sequentially through multiple layer types during production. A balanced WIP generally means that the WIP is well distributed over the different layer types. Therefore, the production targets are used to set a certain amount of throughput per layer type to ensure a well-balanced WIP.

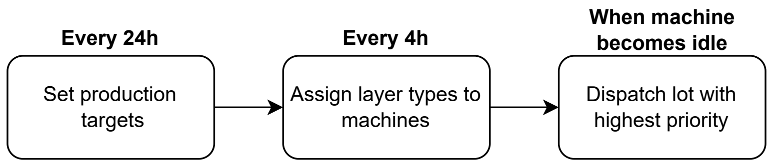


Figure 4: Graphical overview of the steps involved in the current production control.

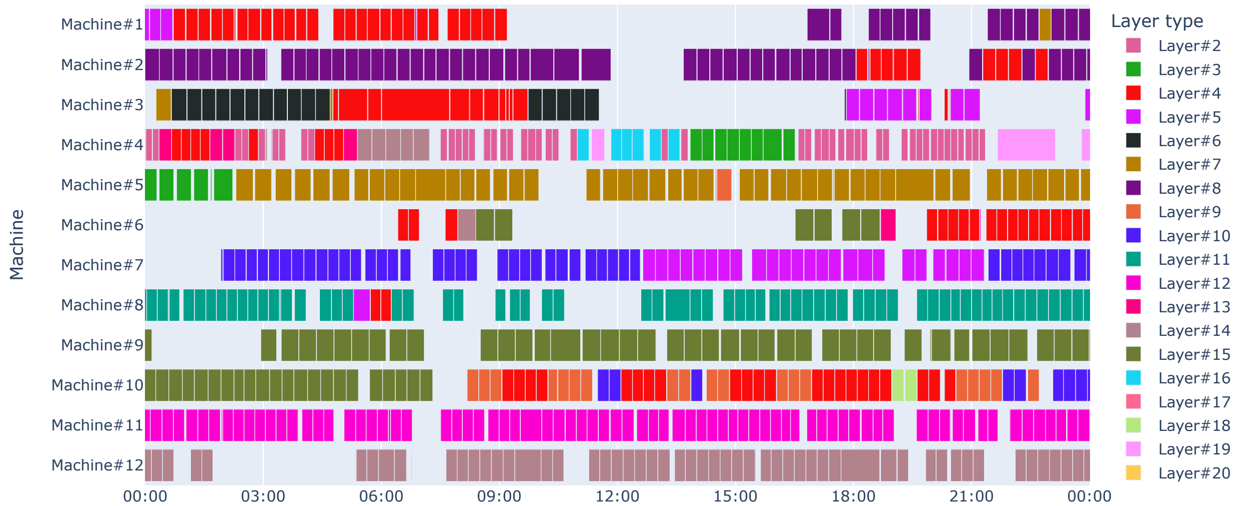
Every four hours, a layer type is assigned to each machine based on the progress towards the production targets of each layer type, e.g. layer types which are further behind schedule get higher priorities. Lots which are in the same layer type, even if the product type is different, have many similarities in processing characteristics such as machine eligibilities and machine speeds. Hence, it makes sense to use the layer type to group the available lots for the machine allocation. It is preferred to assign each layer to a maximum of two machines at the same time to prevent reticle conflicts. When the machine becomes idle, the first eligible lot based on priority is dispatched next. Hot lots get the highest priority, followed by lots of the assigned layer type which uses the same reticle as currently loaded on the machine, followed by lots of the assigned layer type which uses a different reticle. Within these three groups, the lots are prioritized by earliest operational due date first. In case the queue of an assigned layer type is empty before the end of the four hour window, a different layer type is assigned to that machine. Also, when a machine goes down or becomes available again, the layer type to machine assignment is reevaluated. As mentioned before, the current production control is executed by humans and is therefore not always executed exactly according to the described heuristic.

3 RESULTS

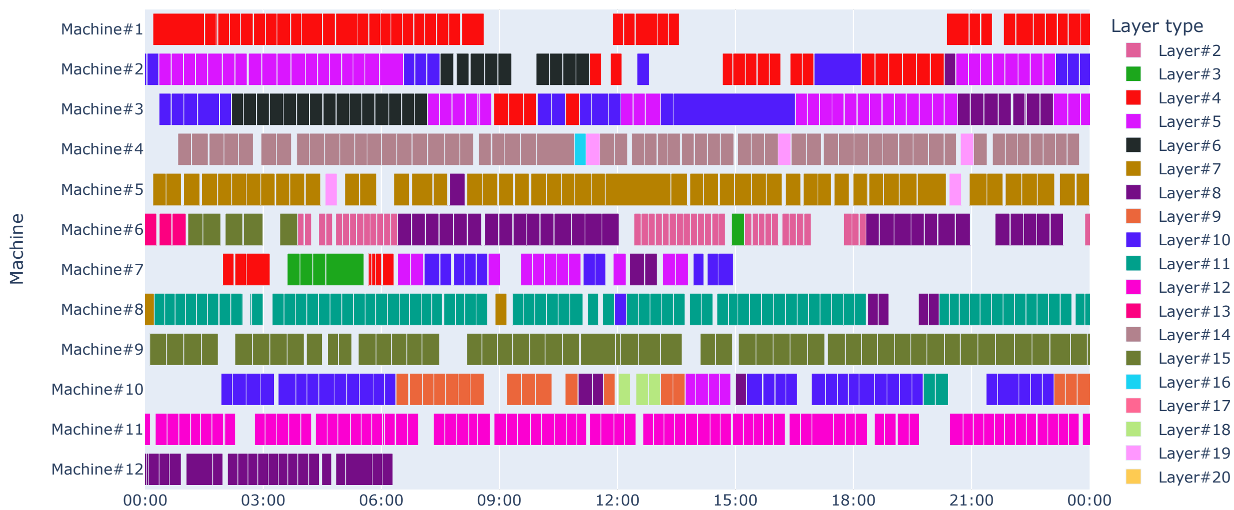
The length of the experiment is 30 days and the number of replications is set to 300. The realized production schedules from the simulations are compared to the single realization from the real-world production from the same time frame. The simulations are performed on a computer with an Intel Core i5-8600K processor running at 3.60 GHz and 16 GB of RAM memory. On this computer, a simulation run of 30 days only takes 2 seconds, i.e., 300 replications takes 10 minutes. This time is excluding the collection and reading in of the production data to build the model, since this is strongly dependent on the used MES and data structures.

Two realized production schedules, one from the real-world and one replication from the simulation model, are depicted in Figure 5. Each bar represents the production of one lot and the color of the bar represents the layer type of the lot. The length of the bar is determined by the stochastic processing time of that lot-machine combination. The non-productive times are caused by either a stochastic down time or stochastic setup time. The layer types are clearly grouped and each layer type is produced on a limited set of machines at the same time. Visually comparing both realizations, we see many similarities in the allocation of layer types to machines. However, we also observe the stochastic nature of the production environment, with many differences in random down times of machines and also slight differences in the layer type allocations.

In the remainder of this section, we will verify the accuracy of the simulation model on different objectives. We will start with analyzing the throughput. After that, we will consider the layer and reticle changes, these two factors determine the behaviour of setup times. Next we will consider the lateness and finally we analyze the accuracy using the semiconductor standards of the overall equipment efficiency (OEE).



(a) Real world



(b) Simulation model

Figure 5: Realized production schedules.

3.1 Throughput

Throughput is one the most important objectives to optimize in a wafer fab. Therefore, the accuracy of the simulation model in terms of throughput is of high importance. The wafers produced per day and queue length at the end of each day of the first 20 replications of the experiment are plotted in Figure 6. It can be seen that the throughput and queue length of the simulation model very accurately matches the throughput of the real-world observation. The stochastic behaviour of the real-world and single replications from the simulation also seem to match well. The mean and standard deviation of the wafers produced per day is respectively 0.15% and 6.60% lower in the simulation model compared to the real-world.

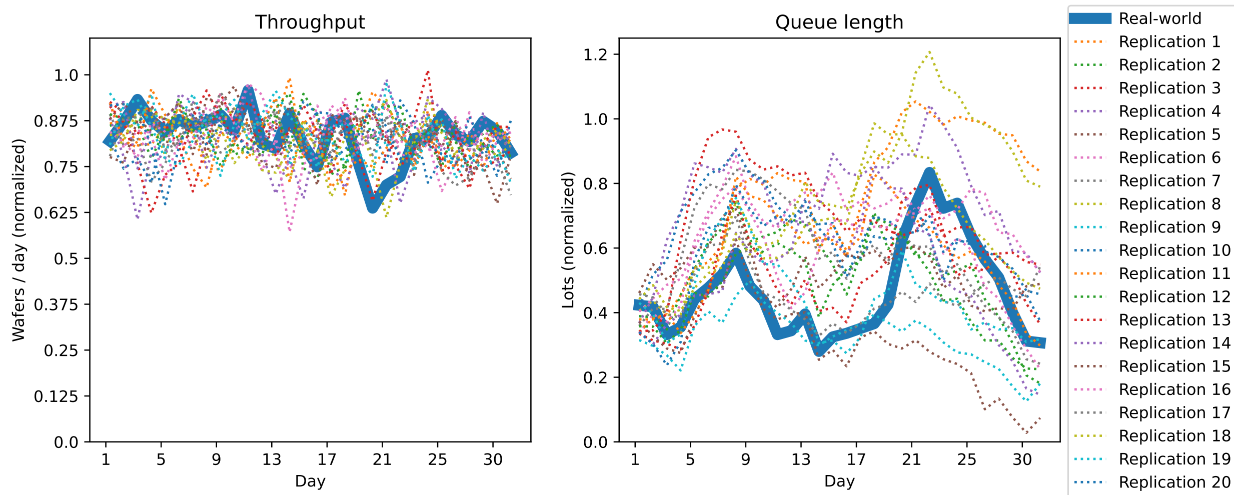


Figure 6: Comparison of the throughput (left) and resulting queue length (right) of all replications of the simulation model (dashed colored lines) with the real-world (solid blue line).

3.2 Reticle and Layer Changes

Reticle and layer changes influence the setup times, as explained in Section 2.2.4. A reticle change happens when the next lot is from the same layer type but needs a different reticle. A layer change happens when the next lot is from a different layer type than the previous lot. The reticle and layer changes are considered to validate how well the current production control is captured in the dispatching heuristic of the simulation model. Figure 7 compares the single real-world observation with the first 20 replications of the simulation model. The mean and standard deviation of the reticle changes in the simulation model are respectively, 49.28% lower and 13.97% higher than the single real-world observation. For the layer changes this is 25.98% lower and 5.73% higher in the simulation model compared to the real-world. This indicates that the simulation model groups the lots more by reticles and layer types than in the real world. Grouping the lots by reticles and layers is preferred behavior of the heuristic and the simulated production control always follows this heuristic precisely. In contrast, the real-world workflow of the operators can sometimes differ slightly from this heuristic. Choosing a lot from the queue and loading it onto the machine is a manual operation and different operators tend to work differently. For instance, operators do not always adhere to this heuristic to ease their personal work flow. Operators sometimes pick another lot which is located closer to them in the fab or a lot which does not require a setup on the machine. This explains the deviation between the observed and simulated reticle and layer changes. However, the impact of this deviation on certain performance measures such as throughput is relatively small.

3.3 Overall Equipment Efficiency (OEE)

Another way to measure throughput performance is with the SEMI standards (North American Metrics Committee 2000), which are international standards for the semiconductor industry. The throughput can be analyzed with the performance efficiency, which is the multiplication of the operational efficiency and the rate efficiency. The operational efficiency is the fraction of the production time divided by the equipment uptime. If no idle time between jobs exist, except for machine downs, this value will be equal to 1. The rate

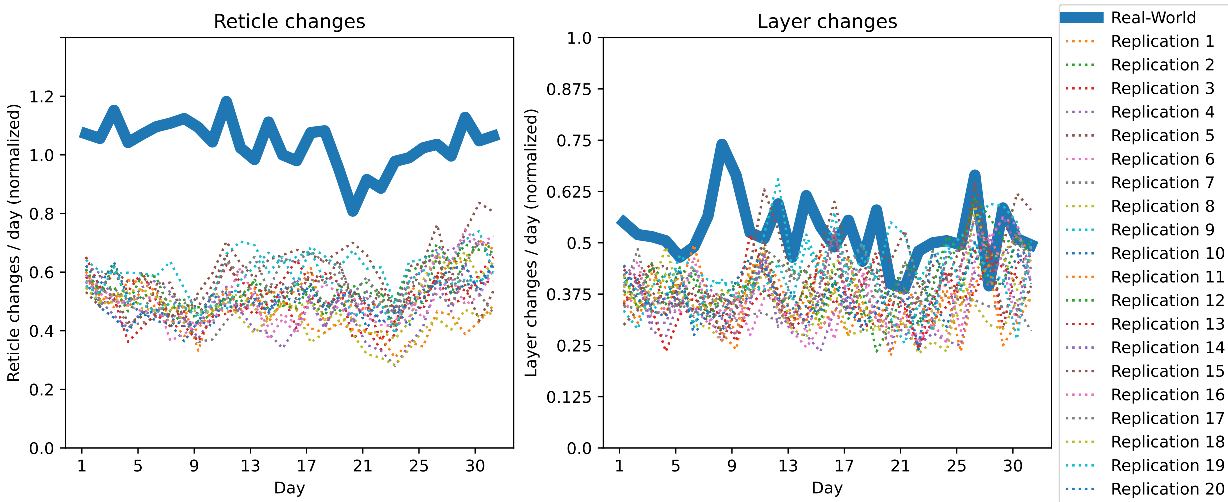


Figure 7: Comparison of the throughput (left) and resulting queue length (right) of all replications of the simulation model (dashed colored lines) with the real-world (solid blue line).

efficiency is the fraction of the fastest possible processing time of the jobs divided by the actual production time. If all jobs are produced by the fastest machine, this value will be equal to 1. The normalized values of these measures are depicted in Figure 8. It can be seen that the performance efficiency is accurate, although the individual components of operational efficiency and rate efficiency slightly deviate.

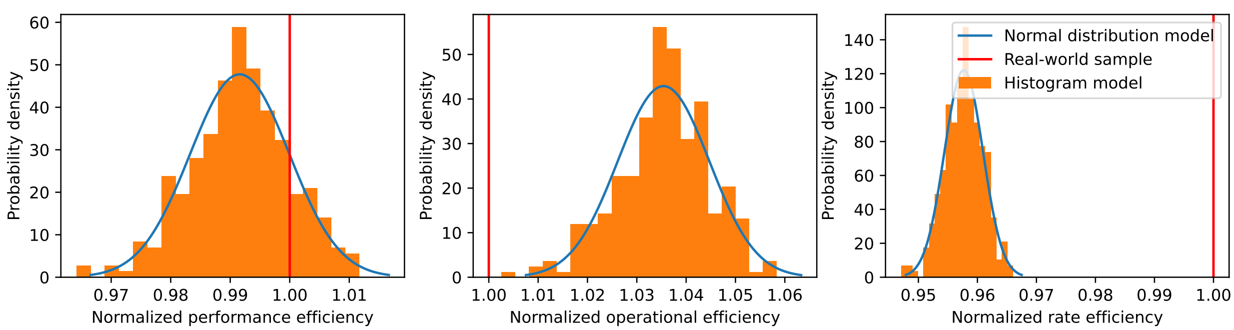


Figure 8: Comparison of the real-world value and the distribution from all the replications of the simulation model in terms of the performance, operational and rate efficiency. Values are normalized to the real-world value.

3.4 Lateness

Another important objective to optimize in the fab besides throughput, is the lateness. A lot which is in the photolithography area yet-to-be-processed is referred to as a job (for the scheduling problem). Each job has an operational due date which indicates when the job should be finished. In order to maximize the on-time delivery performance in the wafer fab, it is essential to keep the wafers on track with their operational due dates. Both earliness and tardiness, with respect to their operational due dates, are penalized. Since lateness combines both earliness and tardiness, this is chosen as an objective. However, the absolute lateness does not differentiate the objective of sequences of jobs which are all being early or all being tardy (the reader is referred to Sun et al. (1999) for an elaborate explanation of this). We want to penalize jobs more if they are further away from the due date. Therefore, the cumulative squared lateness is chosen as objective, which should be minimized by production control methods:

$$\sum_{j \in J} (C_j - d_j)^2 \tag{3}$$

where, J is the set of jobs, C_j is the completion time of job j and d_j is operational due date of job j. This objective is used to quantify the performance in terms of due dates of production control methods, including novel methods in future work. Therefore, we analyze the accuracy of this performance measure. The cumulative squared lateness together with the separate components of the cumulative squared earliness and tardiness are shown in Figure 9. All values are normalized to the observed value of the real-world sample. The results indicate that the simulation model performs slightly better on the total squared lateness objective than the real world, which is mainly caused by the deviating earliness. This is most likely caused by the fact that shop-floor operators do not always exactly follow the priorities as captured in the dispatching heuristic.

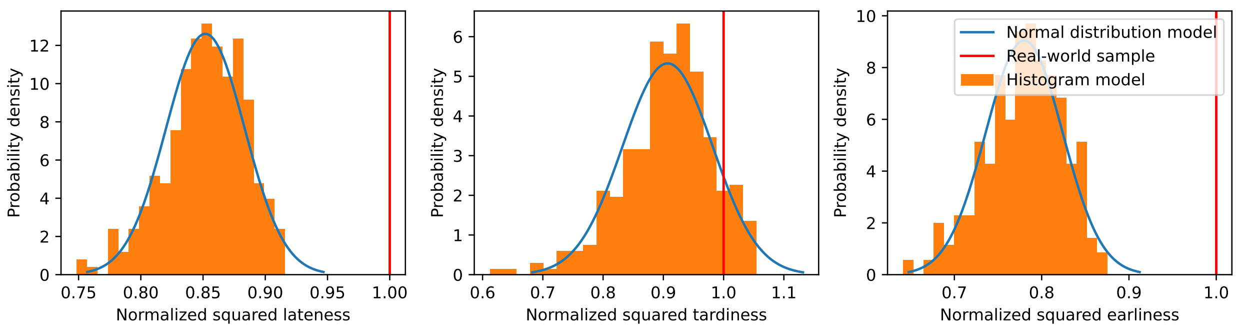


Figure 9: Comparison of the real-world value and the distribution from all the replications of the simulation model in terms of the squared lateness, tardiness and earliness. Values are normalized to the real-world value.

4 CONCLUSIONS AND FUTURE WORK

In this work, we presented a digital twin of the photolithography area which can be used to accurately validate the potential benefit of new production control methods in a stochastic and dynamic environment. The digital twin is a discrete-event simulation model based on real-world data and captures the arrival of jobs, stochastic processing times, stochastic setup times, machine eligibilities and the stochastic machine downs. A real-world case study at Nexperia is presented, but the potential application extends to many other wafer fabs by using the production data and dispatching logic of the corresponding wafer fab.

The results show that the simulation model mimics the real-world production with reasonable accuracy, although there are slight deviations in lateness and performance efficiency. However, the purpose of this simulation model is a test bed to quantify potential benefits of new production control methods. This analysis is done by using the new production control methods in the simulation model and compare it to the current practice of dispatching in the same simulation model, rather than comparing it to the real-world production. Thus, slight deviations will not influence the relative performance and it suffices to show that the simulation model reasonably captures the stochastic and dynamic behaviour.

In future work, we will develop new production control solutions and validate the potential benefit of these solutions in the proposed digital twin. Furthermore, it would be interesting to explore the possibility of the application of this digital twin to other areas than photolithography, other wafer fabs and possibly even other industries.

REFERENCES

Adan, J. and Deenen, P. C. 2021. "C# Simulation Library". https://github.com/JelleAdan/CSSL, accessed 23rd July 2021.

Bitar, A., S. Dauzère-Pérès, C. Yugma, and R. Roussel. 2016. "A Memetic Algorithm to Solve an Unrelated Parallel Machine Scheduling Problem with Auxiliary Resources in Semiconductor Manufacturing". *Journal of Scheduling* 19:367–376.

Cakici, E., and S. J. Mason. 2007, 4. "Parallel Machine Scheduling Subject to Auxiliary Resource Constraints". *Production Planning and Control* 18:217–225.

Doleschal, D., G. Weigert, A. Klemmt, and F. Lehmann. 2013. "Advanced Secondary Resource Control in Semiconductor Lithography Areas: From Theory to Practice". In *Proceedings of the 2013 Winter Simulations Conference*, edited by Raghu Pasupathy, Seong–He Kim, Andreas Tolk, Raymond R. Hill, and Michael E. Kuhl, 3879–3890. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Geiger, C. D., R. Uzsoy, and H. Aytu. 2006. "Rapid Modeling and Discovery of Priority Dispatching Rules: an Autonomous Learning Approach". *Journal of Scheduling* 9:7–34.

Gupta, A. K., and A. I. Sivakumar. 2006. "Pareto Control in Multi-Objective Dynamic Scheduling of a Stepper Machine in Semiconductor Wafer Fabrication". In *Proceedings of the 2006 Winter Simulation Conference*, edited by L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 1749–1756. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ham, A. 2018, 2. "Scheduling of Dual Resource Constrained Lithography Production: Using CP and MIP/CP". *IEEE Transactions on Semiconductor Manufacturing* 31:52–61.

Ham, A. M., and M. Cho. 2015, 8. "A Practical Two-Phase Approach to Scheduling of Photolithography Production". *IEEE Transactions on Semiconductor Manufacturing* 28:367–373.

Janssen, T. 2019. *Optimization in the Photolithography Bay: Scheduling and the Traveling Salesman Problem.* Ph. D. thesis, Delft University of Technology, Netherlands. https://doi.org/10.4233/uuid:12961f87-eeff-41b5-8688-df28e0ad9860, accessed 3rd December 2021.

Mönch, L., J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, and O. Rose. 2011, 12. "A Survey of Problems, Solution Techniques, and Future Challenges in Scheduling Semiconductor Manufacturing Operations". *Journal of Scheduling* 14:583–599.

North American Metrics Committee, S. 2000. *SEMI E79-0200 Standard for Definition and Measurement of Equipment Productivity*. Mountain View, CA: Semiconductor Equipment and Material International (SEMI).

Sarin, S. C., A. Varadarajan, and L. Wang. 2011. "A Survey of Dispatching Rules for Operational Control in Wafer Fabrication". *Production Planning and Control* 22:4–24.

Sun, X., J. S. Noble, and C. M. Klein. 1999. "Single-machine Scheduling with Sequence Dependent Setup to Minimize Total Weighted Squared Tardiness". *IIE Transactions* 31(2):113–124.

AUTHOR BIOGRAPHIES

PATRICK C. DEENEN is a doctoral candidate in the Department of Industrial Engineering of the Eindhoven University of Technology and a Sr. Business Process Analyst at Nexperia. His current research interests are in the area of modeling, control and optimization of manufacturing systems. His email address is patrickdeenen@hotmail.com.

RICK A.M. ADRIAENSEN has obtained his master's degree in the Department of Mechanical Engineering of the Eindhoven University of Technology. During the course of this work, he was a graduate intern at Nexperia. Currently, he is a self-employed Process Engineer and focuses on automated planning and scheduling. His current research interests are production scheduling, discrete-event simulation and manufacturing process optimization. His email address is rick.adriaensen@gmail.com.

JOHN W. FOWLER is the Motorola Professor of Supply Chain Management and recently served as Chair of the Supply Chain Management department in the W.P. Carey School of Business at Arizona State University. His research interests include discrete event simulation, deterministic scheduling, multi-criteria decision making, and applied operations research with applications in semiconductor manufacturing and healthcare. He has published over 130 journal articles and over 100 conference papers. He was the Program Chair for the 2002 and 2008 *Industrial Engineering Research Conferences*, Program Chair for the 2008 *Winter Simulation Conference (WSC)*, and Program Co-Chair for the 2012 *INFORMS National Meeting*. He was the founding Editor-in-Chief of *IISE Transactions on Healthcare Systems Engineering* and currently serves as a Healthcare Operations Management Departmental Editor. He is also an Editor of the *Journal of Simulation* and Associate Editor of *IEEE Transactions on Semiconductor Manufacturing* and the *Journal of Scheduling*. He is a Fellow of the Institute of Industrial and Systems Engineers (IISE) and INFORMS. He served as the Hi PIISE Vice President for Continuing Education, is a former INFORMS Vice President, and served on the WSC Board of Directors. His email address is john.fowler@asu.edu.