Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

SIMULATION-BASED EVALUATION OF LOT RELEASE POLICIES IN A POWER SEMICONDUCTOR FACILITY – A CASE STUDY

Henriette Allgeier
Christian Flechsig
Jacob Lohmer
Rainer Lasch

Chair of Business Management, esp. Logistics
Technische Universität Dresden
Münchner Platz 3
Dresden, 01062, GERMANY

Germar Schneider
Benjamin Zettler

Infineon Technologies Dresden
GmbH & Co. KG
Königsbrücker Straße 180
Dresden, 01099, GERMANY

ABSTRACT

Lot release policies, i.e., the decision which lots to start in production, in what quantity and at what time, have a significant influence on fab performance. Recent research focused on closed-loop policies. However, most studies only demonstrated the feasibility in settings with low-mix or low-volume simulation testbeds. In this paper, we focus on a real-world pre-assembly facility in a high-volume and high-mix semiconductor wafer fab. We conduct an in-depth, deterministic discrete-event simulation in two stages, using real production data and demands. First, we test two existing open-loop lot release policies (random and constant release) against a simple closed-loop release policy. Significant improvements in on-time delivery, bottleneck utilization, and throughput are notable. Second, we compare three closed-loop release policies and indicate which policy provides the best results for certain KPIs like enhanced on-time delivery or reduced tardiness.

1 INTRODUCTION

The production of semiconductors based on silicon technologies is one of the most complex manufacturing processes. The reasons being multiple products with a fast-changing product mix, short product life cycles, re-entrant product flows with routes of several hundred process steps, and the usage of highly-sophisticated and expensive machinery ("tools"). Semiconductor companies usually scale according to Moore's law, i.e., they follow a More-Moore strategy and aim to maximize the output of only a few products. In contrast, organizations that rather strive for diversification through a high product variety, pursue the More-than-Moore strategy (Zhang et al. 2006). These companies face the challenge of producing hundreds of different products at the same time, each with varying customer due dates and therefore have to operate very flexibly (Fowler et al. 2002; Keil et al. 2019; Mönch et al. 2013).

The semiconductor manufacturing process is divided into two stages: Within the highly automated frontend manufacturing, semiconductor devices are produced on silicon wafers. In the following backend, which is often labor-intensive and geographically separated from the frontend manufacturing, the wafers are cut into pieces, packaged, and shipped to customers. The backend usually consists of pre-assembly, assembly, and a test facility (Potoradi et al. 2002; Sivakumar 1999).

Our case company Infineon Technologies Dresden (IFD) has automated its 200 mm (8") customer-specific logic fabrication over the years to one of the highest automated manufacturing lines in the world. The facility can also manage a high product mix, including complex logistical activities (Heinrich et al. 2008). In the year 2011, the company started the first worldwide production of power semiconductors on 300 mm (12") wafers, which runs on a higher product mix in parallel to the 200 mm production line.

Contrary to typical semiconductor manufacturing, the pre-assembly of IFD is characterized by a high degree of automation and short cycle times, being the world´s first fully automated 300 mm pre-assembly production line. Therefore, it is located at the frontend site at IFD. Due to increasing global competition and customer orientation, on-time delivery (OTD) is considered a critical success factor in the semiconductor industry (Kim and Lim 2012). Hence, companies seek for due date compliance by minimizing mean cycle time and maximizing throughput (Kim et al. 2008).

The mentioned characteristics and challenges of the semiconductor industry lead to complex production planning and control, which requires the implementation of dynamic workload control policies, defined as the combination of lot release (also referred to as "order release") with dispatching strategies. Lot release includes a decision at the shop floor that determines the timing, order, and quantity of lots to be released into production. Dispatching is a decision at the individual tool groups that determines which lot in the waiting queue should be processed next on which tool (Fowler et al. 2002; Kim et al. 2001).

Motivated by a recent research project in a consortium of leading semiconductor companies, this paper focuses on the development of an optimized lot release strategy for short-term planning for the high-volume, high-mix pre-assembly facility of the 300 mm wafer fab of IFD. The pre-assembly includes different steps of wafer backside processing as well as the separation of individual chips ("dies") on framed foils through specific sawing processes (Weigert et al. 2009). A recent literature review on release strategies revealed that previous research mainly considered theoretical models based on low-mix scenarios; an exception is the case study of Qi et al. (2009), consisting of 511 machines and 37 part types (Lohmer et al. 2020). However, findings from these studies cannot be easily transferred to a real-world high-volume, high-mix semiconductor fab. Although OTD has a significant impact on performance in semiconductor manufacturing, scientific literature regarding this topic is scarce. Related articles often follow a one-factor-at-a-time approach and focus on cycle time, throughput, and the level of work in process (WIP), neglecting important customer-related performance indicators (KPIs), like due-date compliance and tardiness (Singh and Mathirajan 2018).

This paper presents a simulation study divided into two stages, using real data of a high-volume, high-mix fab to evaluate the performance of several open-loop and closed-loop release policies. The first stage compares two open-loop policies with a closed-loop policy. The second stage presents a full-factorial simulation study by comparing three closed-loop release policies. Besides the KPIs that are referred to most in scientific literature, namely cycle time, standard deviation of cycle time, throughput, WIP, and bottleneck utilization, we further consider OTD, overall tardiness as well as number and tardiness of delayed lots. Hence, we enable companies to adopt their lot release policy according to preferred criteria. Based on a discrete-event simulation approach, this paper contributes to academia by answering the following two research questions (RQ):

RQ1: Which lot release policy is promising for a high-volume, high-mix pre-assembly in the semiconductor industry?

RQ2: Which managerial implications can be derived from the simulation study for high-volume, high-mix facilities in the semiconductor industry?

The remainder of this paper is structured as follows. Section 2 explains the current lot release policies of IFD and briefly reviews the related literature. Sections 3 includes the applied simulation model. The results of our experiments are presented in Section 4 and discussed in Section 5. We conclude our paper with Section 6 and give recommendations for future research.

2 RELATED WORK ON LOT RELEASE POLICIES

Scientific research on lot release strategies dates back to the late 1980s (Glassey and Resende 1988; Wein 1988), demonstrating its greater impact on the overall performance of the system when compared with dispatching (Mönch et al. 2013; Singh and Mathirajan 2018). Release policies can be distinguished in open-loop (OLRP) and closed-loop (CLRP). With OLRPs, lots are released in a specific time interval based on static and exogenous information, like demand, neglecting the current fab situation. In contrast, CLRPs are dynamic. They include real-time shop floor information to determine the optimal release time of the lots

and usually outperform OLRPs (Li et al. 2015). In this section, we describe the OLRP in use at IFD (Section 2.1). Furthermore, we briefly explain related literature on specific CLRPs that are important for our simulation study (Section 2.2). We focus on lot release policies and do not investigate the influence of different dispatching rules but use the currently implemented rules at IFD.

2.1 Open-loop Release Policies and Current Practice of the Case Company

An OLRP appears to be the most obvious way to maintain a certain average throughput rate (Wein 1988). Although resulting in non-optimal outcomes, such policies are widely used in practice (Glassey and Resende 1988; Qi et al. 2009). Scientific literature often compares OLRP to CLRP to demonstrate the superiority of CLRPs (Singh and Mathirajan 2018).

The company in this paper initially used a random (*RAND*) OLRP. Based on the first results of the optimization project, the policy was adapted to a constant (*CONST*) release strategy. *RAND* releases lots into the system immediately (randomly) upon arrival at the beginning of each day or shift. *CONST* (also referred to as uniform, *UNIF*) is the most popular OLRP (Qi et al. 2009; Singh and Mathirajan 2018) and releases lots into the system at a constant rate and time interval (Glassey and Resende 1988; Wein 1988). In their work, Sivakumar and Chong (2001) demonstrate the superiority of *CONST* over *RAND*, concerning throughput, cycle time, and standard deviation of cycle time. Due to the queuing theory, constant lot release reduces the variability of the production system, thus improving its performance. Therefore, *CONST* is assumed to be the best OLRP.

2.2 Closed-loop Release Policies

Since the late 1980s, CLRPs have received increasing interest of scholars, resulting in multiple new lot release strategies. Constant work in progress (*CONWIP*) (Spearman et al. 1990), workload regulation (*WR*) (Wein 1988), and starvation avoidance (*SA*) (Glassey and Resende 1988) represent the most recognized CLRP in academia (Lohmer et al. 2020). *SA* is suitable for facilities with long cycle times, as it only considers lots in a certain time interval (Glassey and Resende 1988). Singh and Mathirajan (2018) compared several CLRPs. They show that *CONWIP*, constant load (*CONLOAD*), *WR*, and their newly developed constant workload (*CONSTWL*) perform best, concerning cycle time and WIP under a defined throughput. *CONLOAD* represents an extension of *WR* and considers the distribution of workload for the bottleneck work center over time instead of the total amount of work for the bottleneck. A new lot is released into the fab if the current bottleneck load plus the workload implied by the new lot is less than a given threshold (Rose 1999). However, *CONLOAD* requires much product-specific information, which makes it suitable only for fabs running on a low product mix. The pre-assembly facility, in our case, is characterized by short cycle times and a high product mix. Thus, *SA* and *CONLOAD* are not regarded in this simulation study. CLRPs considered suitable and easy to implement into the case company are briefly explained in the following:

CONWIP. *CONWIP* is designed to maintain a constant overall WIP level. A new lot is released into the fab only, according to its priority, if the WIP level falls below a given threshold (Spearman et al. 1990). Following Little's law on the relation of average cycle time and average WIP level, limiting the WIP results in shorter cycle times of the lots. *CONWIP* only considers the overall WIP level, neglecting the distribution of workload on specific tools (Singh and Mathirajan 2018). Thus, it appears suitable, especially for balanced fabs running in a steady-state (Rose 1999).

WR. Compared with *CONWIP*, *WR* allows for a more accurate workload regulation as it considers the specific workload of a single lot for the bottleneck (Mönch et al. 2013; Rose 1999). A new lot is released into the fab whenever the current workload plus the total amount of process time of the lot for a given bottleneck station falls below a given threshold (Wein 1988). *WR* refers only to the total bottleneck process time, neglecting its interrelation with the total amount of cycle time (Rose 1999).

CONSTWL. While *WR* assumes a constant deterministic bottleneck, *CONSTWL* considers the total process time of all lots. A new lot is released into the fab if the overall workload falls below a certain

threshold, which is simulated in advance using the target throughput (Singh and Mathirajan 2018). *CONSTWL* overcomes the disadvantages of bottleneck-based release policies and may be better suited for real-world semiconductor facilities, where multiple and changing bottlenecks exist.

3 SIMULATION MODEL

Using simulation models to compare and evaluate different release policies is a common approach in the literature on semiconductor manufacturing (Mönch et al. 2013; Qi et al. 2009; Singh and Mathirajan 2018). We developed a lot-fine discrete-event simulation (DES) model of the entire pre-assembly facility using the simulation software AnyLogic©. AnyLogic© is a multi-method, object-oriented and Java-based simulation software that offers various simulation techniques like discrete-event simulation, agent-based simulation and system dynamics. AnyLogic© is highly flexible and suitable for a wide range of applications. The discrete-event simulation technique is well suited for modeling complex production systems (Ivanov and Rozhkov 2020; März et al. 2011; Zauner et al. 2007). Therefore, AnyLogic© is used as a discrete-event simulation tool in this case study.

In the following, we first describe the model and its structure (Section 3.1), explain the model parameters (Section 3.2), and then turn to the applied experimental design (Section 3.3).

3.1 Model Description

The simulation model is based on real-world deterministic input data. It contains boundary conditions of the real production system as well as cause-and-effect relationships and restrictions. A reference model using Business Process Model and Notation (BPMN) was created in advance to serve as the basis for the subsequent simulation model. BPMN is a standardized modeling language, which has decisive advantages over other reference models, like a high degree of automation and suitability to model complex processes (Jošt et al. 2016). The simulation model and the underlying BPMN model were validated with experts from the case company. Also, several test runs were performed, and the results were compared with the actual Manufacturing Execution System (MES) to ensure the models' validity.

The pre-assembly of IFD is composed of around ten fully automated work centers (WC), which are connected through multiple different process flows. The WCs comprise dozens of tools, different equipment dedications, and setup states with sequence-dependent setup times. The amount and properties of the WCs depend on the manufacturing process and the specific characteristics of the products. For those reasons, the full factory scheduling problem in the company's pre-assembly is categorized as a complex job shop environment (Mönch et al. 2013). The facility has natural restrictions due to the capacity of the transport system that can contain several hundreds of lots in the production system at the same time. Our simulation study is a snapshot of the pre-assembly, as we focus on a specific production phase. Therefore, we determine the maximum capacity of the transportation system with 700 lots, which represents one level of the capacity limit of the transport system without further capacity enhancements. Furthermore, we assume nine WCs and 60 tools involved (see Figure 1). WC 6 is considered the bottleneck of the pre-assembly facility.

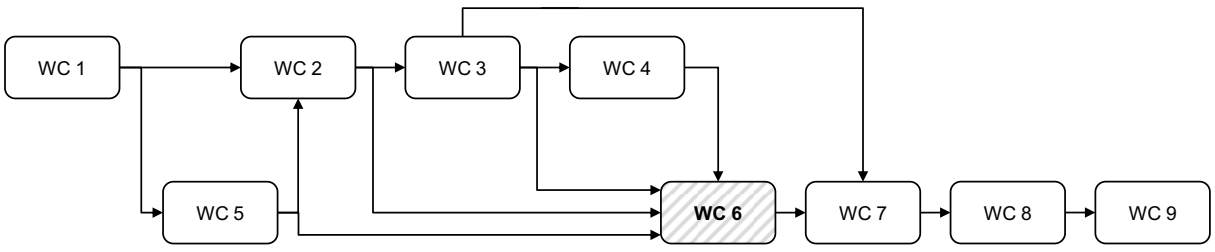

Figure 1: Schematic structure of the pre-assembly facility in the case study with the bottleneck WC6.

3.2 Model Parameters

To illustrate the interaction of different elements in a real-world semiconductor manufacturing system, we differentiate between objectives, input data, and control variables.

Objectives. Several objectives were identified and ranked depending on their priority within the company using input from expert interviews. The overall goal is to achieve a constant workload to avoid variations of the WIP levels through an efficient lot release strategy (Lohmer et al. 2020). The essential measurable objective is to achieve the *highest possible OTD* by *minimizing the total delay (tardiness)* of orders, as well as the *number* and *tardiness of delayed lots*. Besides, capacity utilization in the pre-assembly should be maximized by achieving *high throughput* and *high utilization of the bottleneck*. Furthermore, we aim to *minimize the cycle time* of all lots. The standard deviation of the cycle time shows the dispersion of cycle times for individual lots. *Reducing the standard deviation of the cycle time* has significant advantages, like an enhanced predictability and service level of the production system, resulting in a lower variability of the system and minimized WIP-level variation as well as lower safety stock levels (Qi et al. 2009).

Input Data. The pre-assembly facility is a deterministic production segment that relies on historical planning data. If the simulation is used for medium or long-term planning, suitable stochastic events, such as tool breakdowns, must be defined as stochastic distributions. Simulation studies are often carried out using statistical data on a certain fixed state of the production system. In short-term planning, however, all stochastic elements can be neglected, and deterministic input data can be used. Deterministic means that the system contains no random elements (März et al. 2011). Since all underlying data is known in advance in our setting, we use deterministic input parameters for our model. The input is based on existing capacity planning data and historical data. Therefore, the simulation model does not contain stochastic influences. Considering lot and order data, we include 289 different product types with different process flows, process times, tool dedications, and sequence-dependent setup times.

A full wafer lot consists of 25 wafers. However, smaller lots are permitted, and orders can consist of more than one lot. Each lot has a delivery week assigned, which determines the earliest possible starting time of the lot. Before a lot has been assigned to a delivery week, there is no customer order, and lots are not started into the pre-assembly facility. Cycle time deviation and tardiness measurement are based on the OTD of lots regarding target performance. Lots not delivered within the delivery week (calculated as lot start date plus target cycle time) are delayed. A delivery week contains a specific number of orders that determine the average workload of the production system. As depicted in Figure 1, multiple different process flows exist (that are not detailed due to confidentiality). Transportation time is regarded as an average between the process steps and is based on actual planning data. We have also modeled dispatching rules in production and distinguish between global and local dispatching rules. *First-in, first-out (FIFO)* is considered as the global default, whereas the local rule is *Same Setup*, as lots in the queue are sorted according to the same setup status of the tool. Unscheduled downtime is considered as deterministic input as well, using historical data of the average of the mean time between failure (MTBF) and mean time to repair (MTTR).

Control Variables. We use three control variables in the simulation experiment to assess the performance of the lot release policies under various circumstances. Based on expert panels, we opt for product mix (high, low) and workload (high, medium, low) as well as the lot release policies *RAND*, *CONST*, *CONWIP*, *WR,* and *CONSTWL*.

3.3 Experimental Design

The simulation experiments are carried out in two stages. The shop floor has natural restrictions due to the capacity of the transport system. By using OLRPs, the transport system can be congested, as OLRPs do not consider the current state of the production system. Thus, the transport system works like a natural *CONWIP* threshold, and OLRPs cannot perform better than *CONWIP*, once the capacity limit is reached. Therefore, the selected OLRPs *RAND* and *CONST* are compared with *CONWIP* in Stage 1 of the simulation

experiments to examine the performance of CLRPs compared to OLRPs in general. *CONST* is the policy currently in use at the company, while RAND was used beforehand.

Within the scope of the simulation study, a full-factorial experiment design is created. In Stage 2, the control variables are varied based on different scenarios (parameter configurations). A full-factorial experiment tests all possible combinations of the underlying control variables. This allows gaining insights into the performance of *CONWIP*, *WR*, and *CONSTWL* as well as the influence of different levels of the product mix (low, high) and workload (low, medium, high). As the modeled system represents a high-volume production with a high product-mix, it is necessary to evaluate the impact of a change in product mix and workload. Table 1 summarizes the experimental design, with the three control variables of stage 2 representing the parameters that were varied in the full-factorial experiment, resulting in 18 parameter configurations.

Besides, the number of simulation runs per replication, the length of the warm-up period, and the length of the simulation runs must be determined within the scope of the design of experiments. The results of deterministic simulation models do not change with several simulation runs (Sivakumar and Chong 2001). Representative results are achieved using a simulation run for each parameter configuration. For the evaluation of the simulation results, a comparison of numerical values is sufficient. Stochastic influences only exist for the *RAND* policy. Within the *RAND* setting, 200 repetitions of the runs are performed to achieve a confidence interval of 95% and minimize the effects of a random start distribution. The length of the simulation runs and the warm-up period were determined by several test runs, indicating that a steady state is reached after two weeks (336 hours). In order to achieve a sufficiently long simulation duration, the simulation time is set to 12 weeks (2,016 hours).

Table 1: Experimental design, divided into two stages.

	Control variables	**No. of levels**	**Level**
Stage 1	Lot release policy	3	*RAND, CONST, CONWIP*
Stage 2	Lot release policy	3	*CONWIP, WR, CONSTWL*
	Workload	3	Low, medium, high
	Product mix	2	Low, high

4 EXPERIMENTS AND RESULTS

The objectives (see Section 3.2) were ranked according to the priority of IFD. Scientific literature acknowledges the complex relationship between influencing variables and system performance (Singh and Mathirajan 2018). To compare the performance of different lot release policies, we first select a parameter, on which the other KPIs depend. Usually, the workload is measured by the current WIP level (El-Kilany 2011). An efficient lot release should prevent the congestion of the production system by limiting and balancing the WIP level in production (Bechte 1988). All KPIs change with the parameter configuration, depending on the amount of released material and thus at different WIP levels. For this reason, the WIP level represents the independent variable in this study.

We achieve similar WIP levels by varying the threshold of the lot release strategies. Through a sensitivity analysis, the parameters can be varied in advance to achieve the required output of the simulation model. The inter-arrival time needs to be varied to alter the output of different OLRPs. For CLRPs, the target WIP level can be reached by varying the threshold values. The throughput changes with increasing WIP levels until the WIP level reaches the threshold of 700 lots. A further release above the capacity limit of the system does not change the performance of the production system. The effects on the KPIs are compared by varying the WIP level, ranging from 10 to 700 lots. The threshold of *CONWIP* is varied from 10 to 700 lots in steps of 10 lots. For *WR* and *CONSTWL*, the threshold is a value of time (overall processing time at the bottleneck or in the total system, respectively). It must be varied to reach a certain average WIP

level. For *WR*, the processing time threshold at the bottleneck is adopted from 1,000 to 235,000 minutes. For *CONSTWL*, the processing time threshold of the whole facility ranges from 2,000 to 408,000 minutes.

The number of total simulation runs depends on the number of parameter variations of the sensitivity analysis and the number of parameters (Ragatz and Mabert 1988). In Stage 1, we compare *CONST*, *RAND*, and *CONWIP* at 69 WIP levels and eight KPIs. In Stage 2, we additionally assess *CONWIP*, *WR*, and *CONSTWL* on 18 parameter configurations and eight KPIs. Therefore, we carried out a total amount of 1,380 simulation runs, resulting in 11,040 performance measures. Since the analysis of the extensive results of the experimental design would go beyond the scope of this paper, we will not present the whole raw data.

The results of Stage 1 are presented in Figure 2 graphically. Our results demonstrate the superiority of the CLRP *CONWIP* over the OLRPs *RAND* and *CONST* concerning the critical KPIs in a real production system. *CONWIP* is always superior for WIP levels below 550 lots. For this reason, we consider this policy as the reference for evaluating the performance of *WR* and *CONSTWL* in Stage 2. *CONST* outperforms *RAND,* confirming the findings of Sivakumar and Chong (2001). However, Wein (1988) indicated that the advantages of *CONST* over *RAND* decrease in a higher loaded production system. This result was also observed in our simulation study. Above 550 lots, the policies converge and do not show significant differences, except for OTD and tardiness. Moreover, for high WIP levels, the standard deviation of cycle time and the tardiness of delayed lots for *CONWIP* are worse than *RAND* and *CONST*. *CONWIP* seems to delay lots further that are already late if this enables to optimize the overall system.

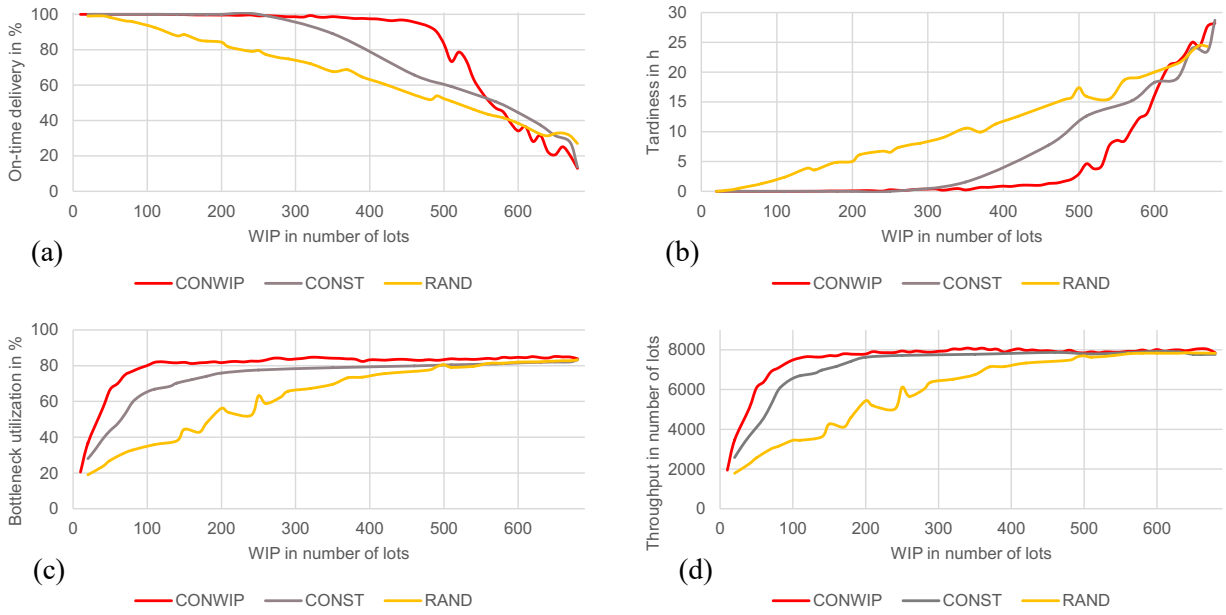

Figure 2: Results of Stage 1 – (a) on-time delivery (OTD), (b) tardiness, (c) bottleneck utilization, and (d) throughput.

For the CLRPs, our simulation study shows less distinct results in Stage 2. To better differ the findings, we use three WIP levels. The low WIP level is defined as 10 to 250 lots, the medium WIP level as 251 to 450 lots, and the high WIP level as 451 to 700 lots. The restriction of the workload through the delivery week leads to a lower reachable WIP level for the parameter configuration with a medium or low workload. For the medium workload, only a low and medium WIP level can be reached. For the low workload, only a low WIP level can be complied. When measuring the performance parameters, delivery reliability of 100% was always obtained at a low workload. Since the differences in the parameter configuration with a low workload are not sufficiently different, we only present the results for a medium and high workload, as shown in Figure 3 and Figure 4. Furthermore, we do not present the results of the parameter configurations

with a low product mix, since the three CLPRs differ even fewer from each other in these settings. Noticeable differences can only be found concerning cycle time and its standard deviation, where *CONWIP* performs slightly better compared to a high-mix setting, but still worse than *WR* and *CONSTWL*.

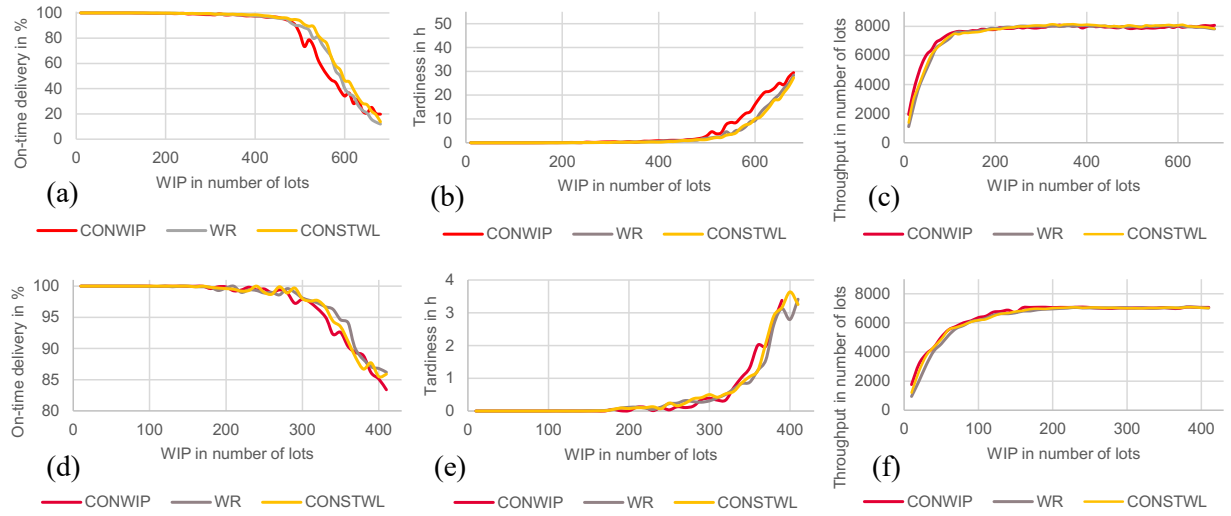

Figure 3: Results of Stage 2 (high-mix setting) – (a)/(d) on-time delivery, (b)/(e) tardiness, and (c)/(f) throughput for high-workload (upper three graphs) as well as medium-workload (lower three graphs).

To better compare the lot release policies, we determine the relative performance deviation in percent. Building on our results of Stage 1, *CONWIP* is considered the basis of the relative deviation, so that *WR* and *CONSTWL* can be compared with it on all performance parameters (KPIs). The relative deviations of the three different WIP levels are presented as an average value for each lot release policy.

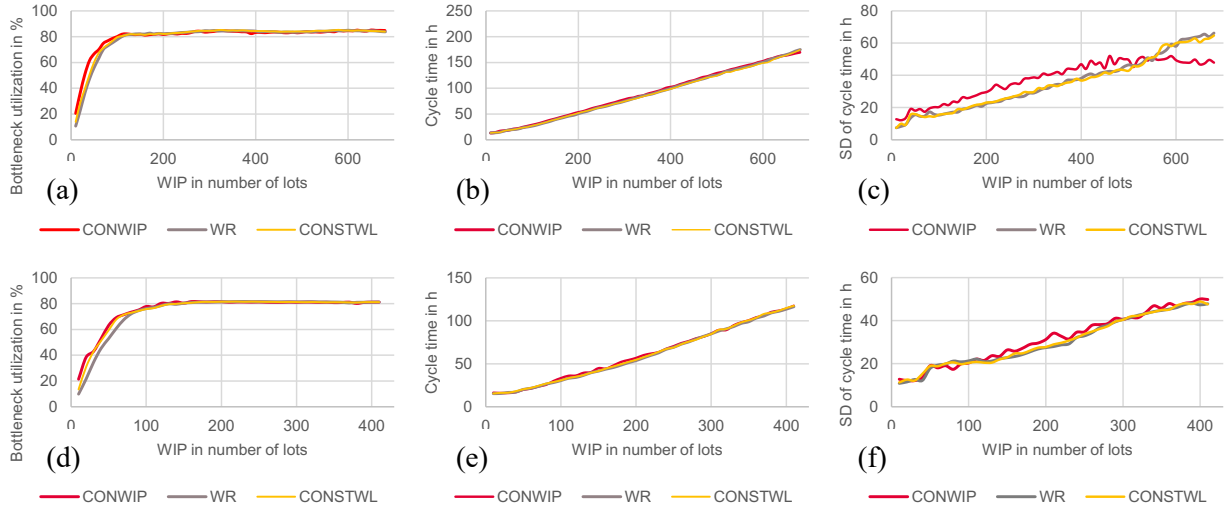

Figure 4: Results of Stage 2 (high-mix setting) – (a)/(d) bottleneck utilization, (b)/(e) cycle time, and (c)/(f) standard deviation (SD) of cycle time (CT) for high-workload (upper three graphs) as well as medium-workload (lower three graphs).

For the high-workload and high-mix scenario, *WR* and *CONSTWL* outperform *CONWIP* regarding OTD and tardiness as well as number and tardiness of delayed lots. In terms of OTD, *WR* works up to 17.2% and *CONSTWL* up to 18.8% better than *CONWIP*. Concerning tardiness, *WR* shows up to 76% and *CONSTWL* up to 68.5% better performance than *CONWIP*. The number of delayed lots is up to 65,2% (*WR*) and up to 59,1 (*CONSTWL*) less than with *CONWIP*. The difference in the tardiness of delayed lots illustrates how well *WR* (up to 56,1%) and *CONSTWL* (up to 43,8%) perform compared to *CONWIP*. Furthermore, both policies show up to 44% (*WR*) and 32% (*CONSTWL*) less cycle time, especially at a low WIP level. In terms of the standard deviation of cycle time, *CONWIP* performs worst up to a WIP level of 550 lots, with *WR* (up to 25.4%) and *CONSTWL* (up to 24.4%) showing better results. The differences of the high-mix, high-workload setting in medium and high WIP levels regarding throughput and bottleneck utilization are not that significant. Here, at a low WIP level, *CONWIP* outperforms *WR* and *CONSTWL*. For medium workloads, the picture is quite similar. *CONWIP* dominates *WR* and *CONSTWL* concerning throughput and bottleneck utilization at a low WIP level but performs worse on OTD and standard deviation of cycle time at a medium WIP level.

While the differences in performance are minor for a low product mix, the differences become more significant at a high product mix. For high-mix settings, the best policy depends on the workload and WIP level. *WR* seems to perform best with increasing WIP levels when considering bottleneck utilization, tardiness, and OTD. However, while *CONWIP* shows the worst results concerning the important KPIs and preferred WIP level of IFD, *WR* and *CONSTWL* often show similar results. *WR* works slightly better overall, due to an improved OTD, and higher bottleneck utilization at the medium WIP level.

5 DISCUSSION

The results of the experiments show that CLRPs work better than OLRPs for most of the considered KPIs in a real production system. The reason why OLRPs perform worse than CLRPs is that they do not refer to the current state of the production system. However, advantages decrease in higher loaded production systems, confirming the findings of Wein (1988). As soon as the production system reaches its capacity limit, the performance of all policies converges, and the performance of *RAND* and *CONST* is similar to *CONWIP*.

Evaluating the different CLRPs appears more challenging. On the one hand, the results overlap at certain WIP levels. On the other hand, none of the release policies is superior in all design points. It seems that the differences between the lot release policies increase with a higher product mix. In our study, changing the product mix does not have an impact on all KPIs. Though, it should be noted that even the low product mix settings in our parameter configurations represent a significantly higher product mix, as IFD follows a More-than Moore strategy. In contrast, companies following a More Moore strategy produce only a few products. *WR* and *CONSTWL* often show similar results and differ from *CONWIP*. This behavior is related to *WR* and *CONSTWL* being based on process times, whereas *CONWIP* is based on the number of lots. The similarity of *WR* and *CONSTWL* results from the process flows of the facility. After the bottleneck (WC6), all process flows merge into a single process flow, which leads to a flow shop environment in the rear area of the facility. Consequently, the performance of the production system strongly depends on the different process flows before the bottleneck. Overall, all three CLRPs show competitive results. However, there are some limitations to the application of the considered CLRPs. As this simulation model is based on deterministic values, no fluctuations in the rear part of the facility were assumed. That is why global release policies, like *CONWIP* and *CONSTWL*, perform well under this setting. When random events occur, which happens regularly in real-world settings, the lots may not be able to leave the facility anymore. The performance of global policies will diminish, as no new lots can be released into production accordingly. Here, an additional evaluation of the different lot release policies with the inclusion of stochastic variables would be useful in future studies. Thus, *WR* is only recommended over *CONWIP* or *CONSTWL*, if there is a constant deterministic bottleneck, which does not frequently change. In this case, *WR* shows its superiority in some customer-related KPIs. To revisit and answer RQ1, we opt for *WR* as the prioritized lot release policy for the crucial KPIs of IFD, especially for a medium WIP level.

At a high WIP level, however, only a few differences in the policies are apparent. The performance of the system depends on congestion: Lots pile up more and cycle time increases, while no higher throughput or capacity utilization can be achieved. So, there is no reason to congest the facility on purpose. The sensitivity analysis of all different WIP levels shows that there is always an optimum WIP level in each scenario, depending on product mix and demand (workload). Since cycle time is increasing nearly linear, the optimum WIP level is reached when throughput, as well as utilization, reach their maximum, while OTD is still maximized. We recommend that a medium WIP level is maintained. It is also important to adopt the correct threshold in order to sustain an appropriate WIP level. The threshold of the different policies is to be defined in advance, whenever product mix and workload changes. Consequently, the managerial implications are derived in the evaluation and choice of the appropriate WIP level and threshold. This also answers our second research question (RQ2).

Existing literature focuses on standardized testbeds frequently, like the MIMAC dataset or the Intel MiniFab (Li et al. 2015; Mönch et al. 2013; Singh and Mathirajan 2018). In contrast, we modeled the real-world system with all production data, constraints, and dedications as well as real demand data in a deterministic setting. Since there was no simulation model of the facility available in advance, we developed a model from scratch, which needed a detailed evaluation and preparation of all the underlying conditions and input data of the real manufacturing system. The result is a profound simulation model of the pre-assembly facility. Unlike other studies, we do not use a one-factor-at-a-time method, but a full factorial doe. We compare two OLRPs and three CLRPs in two stages in a full-factorial design of experiments and 69 parameter variations. The full-factorial design of experiments allows showing the interactions and cause-effect relations between lot release policy, product mix, and workload.

Considering the large variety of semiconductor manufacturing environments, it is evident that there is no universal approach of workload control (respectively lot release) (Mönch et al. 2013). A proper lot release enables semiconductor manufacturers to achieve different production objectives, like short cycle times, high throughput, and a maximized OTD. However, poor planning may still lead to too many lots being released into the production system, which harms various KPIs. Static planning with OLRPs does not take dynamic changes in the production system into account. CLRPs need to incorporate as much information as possible. None of the lot release policies presented here considers conditions of work centers, such as machine failures, in their basic implementation. Therefore, the development of further advanced CLRPs is promising. With this knowledge and depending on the requirements of the production system and the preferred objectives, companies need to select a lot release policy individually and, if necessary, modify and test it in a simulation model before implementing it in a real production system. In addition, an in-depth validation of the model in the company will be necessary for the future. Simulation experts could create different real-world scenarios and use historical data of delivery weeks to determine the quality of the simulation model.

6 CONCLUSION AND FUTURE RESEARCH

In this paper, we describe a two-staged simulation study for a real-world pre-assembly facility in a high-volume and high-mix semiconductor wafer fab at IFD, based on real production data and demands. In Stage 1, we assess *CONWIP* against the OLRPs *CONST* and *RAND*, which have been adopted by IFD in advance. Improvements for *CONWIP*, especially in terms of decisive KPIs, like OTD, tardiness, bottleneck utilization, and throughput, are observed. In Stage 2, we compare *CONWIP* with *WR* and *CONSTWL*, which were derived from literature as the most promising lot release policies and considered suitable for implementation in a real-world semiconductor facility. All CLRPs show competitive results with no policy outperforming the others regarding all KPIs. *WR* and *CONSTWL* often show similar results and differ from *CONWIP*. However, the performance strongly depends on process flows prior to the bottleneck. For IFD, we opt for *WR* as the prioritized lot release policy, as it performs best concerning crucial KPIs, especially for a medium WIP level and a constant deterministic bottleneck. We further recommend maintaining a medium WIP level since OTD decreases, and cycle time increases on a high WIP level. Adopting the correct threshold in order to sustain an appropriate WIP level is also essential.

For future research, we plan to interlink the AnyLogic© simulation model with a mathematical solver for multi-criteria optimization. Further research should also include stochastic influences. Besides, lot release policies that collect as much real-time information about the running state of the facility as possible are promising. Self-learning algorithms can be used to calculate the optimum threshold dynamically. Finally, future research may focus on the impact of several dispatching strategies on lot release policies and therefore evaluate different combinations.

ACKNOWLEDGMENTS

This research was supported in part by the EU project iDev40. The iDev40 project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 783163. The JU receives support from the European Union's Horizon 2020 research and innovation program. It is co-funded by the consortium members, as well as grants from Austria, Germany, Belgium, Italy, Spain, and Romania. The content of this article does not reflect the official opinion of the Joint Undertaking ECSEL. Responsibility for the information and views expressed in the article lies entirely with the authors.

REFERENCES

Bechte, W. 1988. "Theory and practice of load-oriented manufacturing control". *International Journal of Production Research* 26(3):375–395.

El-Kilany, K. S. 2011. "Wafer lot release policies based on the continuous and periodic review of WIP levels". In *2011 IEEE International Conference on Industrial Engineering and Engineering Management*, December 6th–9th, Singapore, Singapore, 1700–1704.

Fowler, J. W., G. L. Hogg, and S. J. Mason. 2002. "Workload Control in the Semiconductor Industry". *Production Planning and Control* 13(7):568–578.

Glassey, C. R. and M. G. C. Resende. 1988. "Closed-loop Job Release Control for VLSI Circuit Manufacturing". *IEEE Transactions on Semiconductor Manufacturing* 1(1):36–46.

Heinrich, H., G. Schneider, F. Heinlein, S. Keil, A. Deutschländer, and R. Lasch. 2008. "Pursuing the Increase of Factory Automation in 200mm Frontend Manufacturing to Manage the Changes Imposed by the Transition from High-Volume Low-Mix to High-Mix Low-Volume Production". In *2008 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, May 5th–7th, Cambridge, Massachusetts, 148–155.

Ivanov, D. and M. Rozhkov. 2020. "Coordination of Production and Ordering Policies under Capacity Disruption and Product Write-off Risk: An Analytical Study with Real-data Based Simulations of a Fast Moving Consumer Goods Company". *Annals of Operations Research* 291(1–2):387–407.

Jošt, G., J. Huber, M. Heričko, and G. Polančič. 2016. "An Empirical Investigation of Intuitive Understandability of Process Diagrams". *Computer Standards & Interfaces* 48:90–111.

Keil, S., F. Lindner, G. Schneider, and T. Jakubowitz. 2019. "A Planning Approach for an Effective Digitalization of Processes in Mature Semiconductor Production Facilities". In *2019 30th Annual SEMI Advanced Semiconductor Manufacturing Conference*, May 6th–9th, Saratoga Springs, New York, 1–6.

Kim, J.-G. and S.-K. Lim. 2012. "Order-lot Pegging for Minimizing Total Tardiness in Semiconductor Wafer Fabrication Process". *Journal of the Operational Research Society* 63(9):1258–1270.

Kim, Y.-D., J.-Y. Bang, K.-Y. An, and S.-K. Lim. 2008. "A Due-Date-Based Algorithm for Lot-Order Assignment in a Semiconductor Wafer Fabrication Facility". *IEEE Transactions on Semiconductor Manufacturing* 21(2):209–216.

Kim, Y.-D., J.-G. Kim, B. Choi, and H.-U. Kim. 2001. "Production Scheduling in a Semiconductor Wafer Fabrication Facility Producing Multiple Product Types with Distinct Due Dates". *IEEE Transactions on Robotics and Automation* 17(5):589–598.

Li, L., Z. Chen, Q. Yu, and N. Xiang. 2015. "Learning-based Release Control of Semiconductor Wafer Fabrication Facilities". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 2965–2973. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lohmer, J., C. Flechsig, R. Lasch, K. Schmidt, B. Zettler, and G. Schneider. 2020. "Order Release Methods in Semiconductor Manufacturing: State-of-the-Art in Science and Lessons from Industry". In *2020 31st Annual SEMI Advanced Semiconductor Manufacturing Conference*, August 24th–26th, held online, in press.

März, L., W. Krug, O. Rose, and G. Weigert. 2011. *Simulation und Optimierung in Produktion und Logistik. Praxisorientierter Leitfaden mit Fallbeispielen*. Berlin, Heidelberg: Springer.

Mönch, L., J. W. Fowler, and S. J. Mason. 2013. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities. Modeling, analysis, and systems*. New York, New York: Springer.

Potoradi, J., O. S. Boon, and S. J. Mason. 2002. "Using Simulation-based Scheduling to Maximize Demand Fulfillment in a Semiconductor Assembly Facility". In *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yücesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes. 1857–1861. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Qi, C., A. I. Sivakumar, and S. B. Gershwin. 2009. "An Efficient New Job Release Control Methodology". *International Journal of Production Research* 47(3):703–731.

Ragatz, G. L. and V. A. Mabert. 1988. "An Evaluation of Order Release Mechanisms in a Job-shop Environment". *Decision Sciences* 19(1):167–189.

Rose, O. 1999. "CONLOAD-a new lot release rule for semiconductor wafer fabs". In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 850–855. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Singh, R. and M. Mathirajan. 2018. "Experimental Investigation for Performance Assessment of Scheduling Policies in Semiconductor Wafer Fabrication – A Simulation Approach". *The International Journal of Advanced Manufacturing Technology* 99(5-8):1503–1520.

Sivakumar, A. I. 1999. "Optimization of Cycle Time and Utilization in Semiconductor Test Manufacturing Using Simulation Based, On-line, Near-real-time Scheduling System". In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 727–735. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sivakumar, A. I. and C. S. Chong. 2001. "A Simulation Based Analysis of Cycle Time Distribution, and Throughput in Semiconductor Backend Manufacturing". *Computers in Industry* 45(1):59–78.

Spearman, M. L., D. L. Woodruff, and W. J. Hopp. 1990. "CONWIP: A Pull Alternative to Kanban". *International Journal of Production Research* 28(5):879–894.

Weigert, G., A. Klemmt, and S. Horn. 2009. "Design and Validation of Heuristic Algorithms for Simulation-based Scheduling of a Semiconductor Backend Facility". *International Journal of Production Research* 47(8):2165–2184.

Wein, L. M. 1988. "Scheduling Semiconductor Wafer Fabrication". *IEEE Transactions on Semiconductor Manufacturing* 1(3):115–130.

Zauner, G., D. Leitner, and F. Breitenecker. 2007. "Modeling Structural - Dynamics Systems in MODELICA/Dymola, MODELICA/Mosilab and AnyLogic". In *Proceedings of the 1st International Workshop on Equation-Based Object-Oriented Languages and Tools*, edited by P. Fritzson, F. Cellier, and C. Nytsch-Geusen, 99–110. Linköping: Linköping University Electronic Press.

Zhang, G. Q., M. Graef, and F. van Roosmalen, F. (2006): "The Rationale and Paradigm of "More than Moore"". In *56th Electronic Components and Technology Conference 2006*, May 30th–June 2nd, San Diego, California, 151–157.

AUTHOR BIOGRAPHIES

HENRIETTE ALLGEIER is a working student at Infineon Technologies Dresden and currently pursuing her M.Sc. in Industrial Engineering & Management at Technische Universität Dresden. Her email address is henriette.allgeier@mailbox.tu-dresden.de

CHRISTIAN FLECHSIG is a research associate and Ph.D. Candidate at the Chair of Business Management, esp. Logistics at Technische Universität Dresden. He holds an M.Sc. in Industrial Engineering & Management at Technische Universität Dresden, received in 2018. His research focuses on the optimization and automation of business processes, with a main focus on the semiconductor industry. His email address is christian.flechsig@tu-dresden.de

JACOB LOHMER is a research associate and Ph.D. Candidate at the Chair of Business Management, esp. Logistics at Technische Universität Dresden. He holds an M.Sc. in Industrial Engineering & Management at Technische Universität Dresden, received in 2018. His research focuses on production planning and scheduling, distributed manufacturing, manufacturing networks, and digital technologies for OM and SCM. His email address is jacob.lohmer@tu-dresden.de

RAINER LASCH is a Full Professor at the Chair of Business Management, esp. Logistics at Technische Universität Dresden., Besides, he is a high-profile research partner of the BMBF and industry, particularly in the areas of benchmarking in logistics, market-oriented process design, and quantitative planning procedures in logistics. His email address is rainer.lasch@tu-dresden.de

GERMAR SCHNEIDER is a Senior Specialist for Factory Integration and Thin Wafer Handling at Infineon Technologies Dresden. He holds a Ph.D. in Analytical Chemistry from the University of Ulm, received in 1995. His email address is germar.schneider@infineon.com

BENJAMIN ZETTLER is an Expert for Production Line Control with a focus on WIP Flow Management and Controlling at Infineon Technologies Dresden. He holds an M.Sc. in Industrial Engineering & Management at Technische Universität Dresden, received in 2017. His email address is benjamin.zettler@infineon.com