Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo and M. Loper, eds.

DATA-DRIVEN PRODUCTION PLANNING FORMULATIONS FOR WAFER FABS: A COMPUTATIONAL STUDY

Tobias Völker
Lars Mönch

Department of Mathematics and Computer Science
University of Hagen
Universitätsstraße 1
Hagen, 58097, GERMANY

ABSTRACT

Cycle times are of order of ten weeks in most semiconductor wafer fabrication facilities (wafer fabs). They have to be explicitly considered in production planning. A nonlinear relation between resource workload and cycle time can be observed. In this paper, we study data-driven (DD) production planning formulations. These formulations are based on a set of system states representing the congestion behavior of the wafer fab with work in process (WIP) and resulting output levels. The effects of different WIP-output relations and additional capacity constraints in the DD models are investigated. Moreover, several methods are proposed to obtain representative sets of system states. The performance of the DD variants is compared with the performance of the allocated clearing function (ACF) model using a scaled-down simulation model of a wafer fab. Simulation results demonstrate that under certain experimental conditions, the DD models lead to similar profit and cost values as the ACF model.

1 INTRODUCTION

Wafer fabs belong to the most complex manufacturing systems (Chien et al. 2011). Integrated circuits are produced layer by layer on wafers, thin discs made from silicon. Hundreds of expensive machines are used in wafer fabs. They are organized in work centers. Lots consisting of up to 25 wafers are the moving entities. Different types of processes, i.e. batch and serial, can be found in wafer fabs. A batch is a group of lots that are processed at the same time on a machine. Sequence-dependent set-up times, auxiliary resources, and tight customer due dates are common process conditions. A large number of products and a diverse product mix are typical for wafer fabs. Up to 800 process steps. i.e. operations, can belong to a route for advanced products (Mönch et al. 2013). Re-entrant process flows are characteristic for wafer fabs, i.e., the same work center is visited up to 40 times by certain lots. Cycle time, the time span between material being released into the wafer fab and its emergence as finished product is of the order of ten weeks in most wafer fabs.

Production planning is an important function in wafer fabs (Mönch et al. 2018). Since cycle times are long, they must be explicitly considered in production planning models. On the one hand, queuing models, discrete-event simulation, and industrial observations show that cycle times increase nonlinearly with resource utilization. On the other hand, the release decisions made by planning models determine utilization. These observations suggest that cycle times should be treated as endogenous to the planning problem. Workload-dependent lead times must be taken into account in production planning models. Lead times are cycle time estimates in planning formations. In the present paper, we study modifications of the DD formulation proposed by Omar et al. (2017). DD models are based on a set of system states representing the congestion behavior of the fab with WIP and resulting output levels. They are an DD alternative to planning models based on nonlinear CF. We show by simulation experiments that the DD formulations can provide comparable performance to the ACF model of Asmundsson et al. (2009) if the set of system states is chosen in an appropriate way.

The paper is organized as follows. In the next section, we describe the problem and discuss related work. Section 3 provides the production planning formulations that are investigated in this paper. This includes several DD model variants. Approaches to determine the required system states are discussed

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

in Section 4. Results of computational experiments are presented in Section 5. Conclusions and future research directions are provided in Section 6.

2 PROBLEM SETTING

2.1 Production Planning for Wafer Fabs

Production planning involves the allocation of available capacity among the process steps of the products to match supply with given demand in some near-optimal manner. Releases into the wafer fab are determined. However, we know from queuing theory (Buzacott and Shanthikumar 1993), discrete-event simulation experiments (Fowler at al. 2015), and industrial observations (Wu 2005) that the mean and variance of the cycle time increase nonlinearly with resource utilization, which, in turn, is determined by the release decisions made by production planning. The observed circularity implies that cycle times are an output of production planning rather than an input. Hence, cycle times are variables to be controlled in planning models, rather than exogenous parameters that must be estimated. Most of the production planning models in the literature are based on lead times, exogenous parameters independent of resource utilization (Voß and Woodruff 2003). This approach leads to computationally tractable models based on linear programming (LP), but fails to represent the congestion of the wafer fab correctly. Only recently, research is initiated that explicitly addresses this circularity. There are several approaches to take into account workload-dependent lead times in production planning models which will be discussed next.

2.2 Discussion of Related Work and Problem Statement

Iterative methods combine LP models with exogenous lead times with simulation, queuing, or scheduling models to update lead times (Missbauer and Uzsoy 2020). However, the convergence behavior of these methods is unclear (Missbauer 2020), and they require time-consuming simulation runs for planning. Nonlinear optimization models based on queueing concepts to represent the cost of congestion form the second class of approaches. Among them CF-based models are popular. A CF estimates the average output of a work center in a planning period as a function of its available workload in that period. While early CF-based models had difficulties to deal with multiple products, the ACF formulation by Asmundsson et al. (2009) addresses this situation. We know from computational studies that this formulation outperforms models with exogenous lead times that are an integer or fractional multiple of the period length (Kacar at al. 2012; Kacar et al. 2013, Kacar and Uzsoy 2015; Kacar et al. 2016). This result carries over to rolling horizon settings (Ziarnetzky et al. 2015; Häussler et al. 2020). An appropriate parameterization of the CFs is required. One obvious limitation of CF-based production planning approaches is that yet no rigorous methodology for estimating CFs from data is known. This constitutes a large obstacle to their widespread adoption in planning models (Gopalswamy and Uzsoy 2019).

Production planning models based on CFs can be considered as a parameterized approach. However, there are DD approaches that make the parameterization effort for CFs to some extent obsolete. Omar et al. (2017) propose an alternative DD production planning model. The DD model represents a planning approach that utilizes system states. System states consider different products and their relations, but take an aggregated view on the resources and process steps of the production system. They provide expected output values for discrete average WIP values of all products. It is assumed that the system is in steady state, i.e., the distributions of WIP and output are constant over time. Omar et al. (2017) and Gopalswamy and Uzsoy (2018) use the Mean Value Analysis (MVA) proposed by Suri and Hildebrant (1984) to determine system states. However, process conditions typical for wafer fabs such as parallel machines or batching are not taken into account in the MVA.

In the present paper, system states are determined by terminating simulation runs. We are interested in identifying sets of system states that are sufficient to obtain high-quality production plans. Moreover, we investigate the effects of different WIP-output relations and additional capacity constraints on the performance of the DD approach. In contrast to Gopalswamy and Uzsoy (2018) where the initial WIP values for all products are 0, we use initial WIP values obtained from simulation runs since this setting is common for the application of production planning models in a rolling horizon setting. This allows a

more realistic performance assessment of DD approaches relative to the performance of the ACF formulation.

3 PRODUCTION PLANNING FORMULATIONS

3.1 ACF Formulation

For the sake of completeness, we start by recalling the ACF model. We assume that the finite planning horizon of length T is divided into discrete periods of equal length. The model is given as follows:

Sets and indices

t: period index
g: product index
G: set of all products
k: work center index
K: set of all work centers
l: operation index
$O(g)$: set of all operations of product g
$O(g,k)$: set of all operations of product g that can be performed on machines of work center k
$K(g,l)$: set of work centers that can be used to perform operation l of product g
n: segment index
$C(k)$: set of indices denoting the linear segments used to approximate the CF for work center k

Decision variables

Y_{gtl}: quantity of product g completing its operation l in period t
Y_{gt}: output of product g in period t from the last operation of its routing
X_{gtl}: quantity of product g starting operation l in period t
W_{gtl}: WIP of product g at operation l at the end of period t
I_{gt}: finished goods inventory (FGI) of product g at the end of period t
B_{gt}: backlog of product g at the end of period t
Z_{gtl}^{k}: fraction of output from work center k allocated to operation l of product g in period t

Parameters

h_{gt}: unit FGI holding cost for product g in period t
b_{gt}: unit backlog cost for product g in period t
ω_{gt}: unit WIP cost for product g in period t
α_{gl}: processing time of operation l of product g
D_{gt}: demand for product g during period t
β_k^n slope of segment n of the CF for work center k
μ_k^n: intercept of segment n of the CF for work center k.

The ACF model can be stated as follows:

$$\min \sum_{g\in G}\sum_{t=1}^{T}\left(\omega_{gt}\sum_{l\in O(g)}W_{gtl}+h_{gt}I_{gt}+b_{gt}B_{gt}\right) \tag{1}$$

subject to

$$W_{g,t-1,l}+X_{gtl}-Y_{gtl}=W_{gtl}, \qquad g\in G, t=1,\dots,T, l\in O(g) \tag{2}$$

$$I_{g,t-1}+Y_{gt}-B_{g,t-1}+B_{gt}-I_{gt}=D_{gt}, \qquad g\in G, t=1,\dots,T \tag{3}$$

$$\alpha_{gl} Y_{gtl} \leq \mu_k^n Z_{gtl}^k + \beta_k^n \alpha_{gl}\left(X_{gtl} + W_{g,t-1,l}\right), \quad g \in G, t = 1, \ldots, T, l \in O(g), k \in K(g,l), n \in C(k) \tag{4}$$

$$\sum_{g \in G} \sum_{l \in O(g,k)} Z_{gtl}^k = 1, \quad t = 1, \ldots, T, k \in K \tag{5}$$

$$W_{gtl}, I_{gt}, B_{gt}, X_{gtl}, Y_{gtl}, Z_{gtl}^k \geq 0, \quad g \in G, t = 1, \ldots, T, l \in O(g), k \in K(g,l). \tag{6}$$

The objective function (1) is the sum of WIP, FGI, and backlog cost over all products and periods. WIP variables and WIP balance constraints (2) are included to compute the WIP cost in the objective function. The FGI material balance at the end of the line is represented by constraint set (3). The CF relates the expected output of each work center in a period to the planned load of the work center in that period in constraints (4). The output allocation among operations is modeled by constraint set (5). The Z_{gtl}^k ~~Z_{gtl}^k~~ variables scale up the available workload of product g at the beginning of period t to approximate the total workload of all products in that period. This yields an upper bound on the output of product g at work center k. We refer to Asmundsson et al. (2009) and Missbauer and Uzsoy (2020) for the details of the ACF model.

3.2 DD Formulations

The model is parameterized by a set R of system states that allow for different output configurations for each period. A single system state is chosen for each period t of the planning horizon through binary decision variables Γ_{rt} with the state index r such that the total cost is minimized. The following additional notation compared to the ACF model is used:

Sets and indices
r: state index
R: set of all system states r

Decision variables
W_{gt}: WIP of product g at the end of period t
Γ_{rt}: binary variable taking on the value 1, if system state r is chosen in period t, and 0 otherwise

Parameters
Q_{gr}: WIP level of product g in system state r
O_{gr}: expected output quantities of product g in system state r.

The basic DD formulation is given as follows:

$$\min \sum_{g \in G} \sum_{t=1}^{T} \left(\omega_{gt} W_{gt} + h_{gt} I_{gt} + b_{gt} B_{gt}\right) \tag{7}$$

subject to

$$W_{g,t-1} + X_{gt} - Y_{gt} = W_{gt}, \quad g \in G, t = 1, \ldots, T \tag{8}$$

$$I_{g,t-1} + Y_{gt} - B_{g,t-1} + B_{gt} - I_{gt} = D_{gt}, \quad g \in G, t = 1, \ldots, T \tag{9}$$

$$\sum_{r \in R} Q_{gr} \Gamma_{rt} = W_{gt}, \quad g \in G, t = 1, \ldots, T \tag{10}$$

$$\sum_{r \in R} O_{gr} \Gamma_{rt} = Y_{gt}, \quad g \in G, t = 1, \ldots, T \tag{11}$$

$$\sum_{r \in R} \Gamma_{rt} = 1, \quad t = 1, \ldots, T \tag{12}$$

$$W_{gt}, I_{gt}, B_{gt}, X_{gt}, Y_{gt} \geq 0,\ \Gamma_{rt} \in \{0,1\}, \quad g \in G, t = 1, \ldots, T, r \in R. \tag{13}$$

The objective function (7) is the same as (1) taking into account $W_{gt} = \sum_{l \in O(g)} W_{gtl}$. Constraints (8) and (9) are the analogues to (2) and (3). Constraints (10) and (11) lead to matching WIP and output values by setting the decision variables to values provided by a system state $r \in R$, i.e., system state r corresponds to the pair (Q_{rt}, O_{rt}). Constraints (12) ensure that exactly one system state is chosen per period. Finally, constraint set (13) models that the decision variables are nonnegative and binary, respectively.

The period length has to be sufficiently large in relation to cycle times to correctly indicate the output levels for the WIP at the end of period t within the same period. With equidistant lot releases over time as determined by the X_{gt} quantities, WIP levels eventually reach a suitable distribution and matching output quantities. This is not the case anymore if cycle times go beyond the period length. While increasing X_{gt} enables system states with higher WIP values and therefore output values, this might decrease output as newly released lots compete for scarce capacity with almost completed lots. We expect that with sufficiently high cycle times the WIP pattern at the beginning of t or the average WIP within a period are more indicative of the expected output levels within period t. To determine output based on the WIP at the beginning of the period, we propose the following two modifications of constraints (10):

$$\sum_{r \in R} Q_{gr} \Gamma_{r1} = W_{g0} + \Delta_g^+ - \Delta_g^-, \qquad g \in G \tag{14}$$

$$\sum_{r \in R} Q_{gr} \Gamma_{rt} = W_{g,t-1}, \qquad g \in G, t = 2, \dots, T. \tag{15}$$

The binary decision variables Γ_{rt} determine the WIP levels $W_{g,t-1}$ at the beginning of each period t instead of those at its end in constraints (14) and (15). As W_{g0} is a parameter and not a decision variable, it cannot be adjusted to match a chosen system state. Instead, we want the model to choose a state with WIP levels as close as possible to W_{g0} to approximate the expected output within the first period. We therefore introduce the additional decision variables $\Delta_g^+, \Delta_g^- \geq 0$ to allow for deviations. Their values are minimized by adding $M(\Delta_g^+ + \Delta_g^-)$ to the objective function (7) where M is a sufficiently large number. For a formulation based on the average WIP values within a period we substitute (10) with

$$\sum_{r \in R} Q_{gr} \Gamma_{rt} = \frac{1}{2}(W_{g,t-1} + W_{gt}), \qquad g \in G, t = 1, \dots, T, \tag{16}$$

such that the average of $W_{g,t-1}$ and W_{gt} corresponds to the WIP pattern of the chosen system state for period t. A different treatment of the first period is not necessary contrary to constraint set (15) since system states can be chosen by adjusting W_{g1}. We will use the term WIP point to differentiate between the relevant period or the average of the WIP for a system state.

The release quantities in the DD model are limited by the WIP differences of system states for consecutive periods plus the corresponding output levels for the same period as defined by constraints (8). System states that lead to a large bottleneck utilization (BNU) have a higher WIP to output ratio which makes them less desirable in terms of WIP cost per unit of output. However, the transition from a low to a high WIP state might lead to congestion as cycle times increase, and it takes a longer time relative to the period length for lots to distribute evenly throughout the production system. This also leads to a higher deviation from the steady state assumption for the system states which makes the predicted output levels less accurate. To mitigate these effects, we add the constraints

$$\sum_{g \in G} \alpha_{gk} X_{gt} \leq m C_k, \qquad k \in K, t = 1, \dots, T \tag{17}$$

to the model that limit the release quantities X_{gt} and thereby an increase in WIP between periods. By multiplying the X_{gt} quantity with the total average processing time a_{gk} for each product g at work center k, we obtain a measure of how much capacity will be claimed at k in the long run with constant release quantities. This value will be bound by a multiple $m \geq 1$ of the available capacities C_k. The parameter m must be chosen carefully, as low values might be infeasible with high cycle time to period length ratios where the WIP to output ratios of the system states will be high as well. Note that a decrease in WIP levels between periods is already limited by the output quantities connected to the system states at that WIP level.

With the basic DD model (7)-(13) and the constraints (14)-(16) we have a total of three model variants that are summarized in Table 1. The additional input constraints (17) lead to three more variants that are abbreviated by DD_W_t_C, DD_W_{t-1}_C, and DD_W_{avg}_C, respectively.

Table 1: DD model variants without constraint set (17).

Abbreviation	Characteristics	Model
DD_W_t	The output quantities for period t are determined by the WIP levels at the end of the period.	(7)-(13)
DD_W_{t-1}	The output quantities for period t are determined by the WIP levels at the beginning of the period.	(7)-(9),(11)-(13), (14), (15)
DD_W_{avg}	The output quantities for period t are determined by the average WIP levels of the period.	(7)-(9),(11)-(13), (16)

The ACF and DD formulations differ in terms of type and number of decision variables and constraints. The ACF formulation is an LP. Instances of this model can be solved efficiently. Nevertheless, the computational burden for large-sized instances can be high due to the large number of decision variables and constraints. Generating high-quality CFs for an accurate model is subject of ongoing research and requires a large amount of simulation time. The DD formulation is a mixed integer linear program (MILP) with binary Γ_{rt} variables. We expect that the performance of the DD formulation depends on $|R|$. The number of required system states at the same average distance between adjacant states in each product dimension grows exponentially with the number of products. At some point, it will be impossible to generate a sufficient number of states or to solve the model with satisfactory accuracy within a reasonable amount of computing time. Contrary to the ACF model, the number of work centers and operations per product do not influence the size of the instances. The number of decision variables and constraints is much smaller compared to the ACF case.

4 CHARACTERIZING AND GATHERING SYSTEM STATES

4.1 Simulation Model

We use the discrete-event simulation model of Kayton et al. (1997) for the experiments. The model represents a scaled-down wafer fab with typical attributes such as reentrant process flows, batch processing, machine breakdowns, and multiple products. Eleven work centers are in the model. The three products have 22, 14, and 14 operations, respectively. A product mix of 3:1:1 is used. Product 1 visits the bottleneck work center six times, product 2 four times while product 3 is processed on the alternative photolithography work center. The batch machines can process between two and four lots. These machines are the main source of variability in addition to the unreliable machines, interrupting the flow of arriving lots for subsequent work centers. The processing times are log-normally distributed. The lots are processed at each work center using the First-in-first-out (FIFO) dispatch rule. Time to failure and time to repair at unreliable work centers follow gamma distributions. An implementation of the model for the simulation engine AutoSched AP is publicly available (Kayton Model 2021).

4.2 Approaches to Determine Appropriate System States

The WIP and output patterns for system states can be derived from queueing theory, simulation, or from data found in shop floor application systems. In the present paper, system states are generated using the Kayton simulation model to obtain measures of its steady state behavior. The sample points are defined by setting fixed release quantities X_g per period within the capacity limit of the model for each product such that $\sum_{g \in G} \alpha_{gk} X_g \leq C_k, k \in K$. Single lots will be released into the simulated system with matching constant inter arrival times. After a warmup period, the simulation model runs for another 365 days to record the average WIP and output per period for each system state.

We apply the sampling methods Grid, Vargrid, and Stochastic. The sampling points are evenly placed in each product dimension for system state sets of type Grid. Sample points beyond output quantities of 80 lots per week for product 1 and 30 lots per week for product 2 and 3 are discarded to

limit the cardinality of the system state set. We generate two sets with a step size of 7 and 5, respectively. The distribution of output and WIP values for product 1 and 2 for a step size of 5 are depicted in Figure 1.

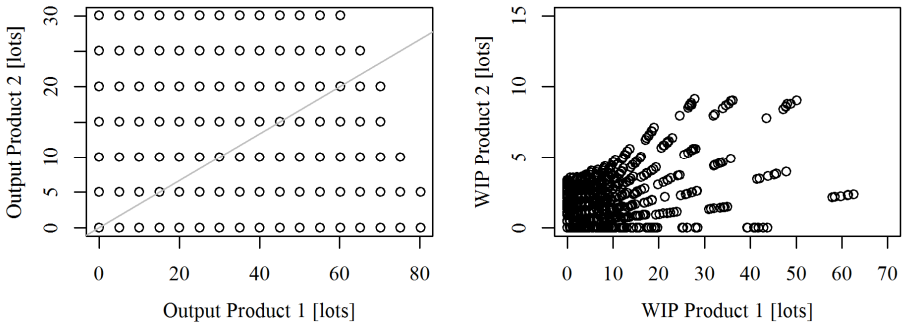

Figure 1: System state set of type Grid with step size 5 (Grid5).

Product 3 is less interesting since it is not processed on the bottleneck work center. Each visible point represents one or more system states. The grey line in the left-hand plot marks the product mix. While there is a good coverage of possible output values, the density in terms of WIP is low beyond values of 20 and 4 for product 1 and 2, respectively. To increase utilization up to the highest levels, large gaps in terms of WIP have to be bridged in subsequent periods.

To increase the density of system states at high utilization and around the product mix, we vary the step size of the grid for sets of type Vargrid. Starting with a base value, the step size is halved for each of three increasingly restrictive conditions leading to four different regions. Each condition $\langle (X_1^{min}, \cdots, X_{[G]}^{min}), \rho^{min}, \Pi^{max} \rangle$ defines minimum values for release quantities $X_g^{min}, g \in G$, expected utilization ρ^{min}, and a maximum deviation from the product mix denoted by Π^{max}. A sample point is within the specified deviation from a given product mix PM, if

$$\frac{1}{\Pi^{max}} \frac{PM_g}{\sum_{p \in G \setminus g} PM_g} \leq \frac{X_g}{\sum_{p \in G \setminus g} X_p} \leq \Pi^{max} \frac{PM_g}{\sum_{p \in G \setminus g} PM_g}, g \in G \tag{18}$$

holds. The conditions $\langle (0,0,0), 0.0, 2.5 \rangle$, $\langle (0,0,0), 0.5, 1.5 \rangle$ and $\langle (0,0,0), 0.7, 1.1 \rangle$ are applied to generate two system state sets with a base step size of 16 and 12. The result of the sampling process and simulation for Vargrid12 is shown in Figure 2.

The density gets higher around the product mix with increasing utilization, while combinations of low output quantities are still covered for both products. We see from the right-hand plot that the WIP levels of system states at high utilization are much closer compared to sets of type Grid. The idea behind state sets of type Stochastic is to obtain states with a distribution that resembles that of the expected demand for all products. First, we define a number of utilization levels $U = \{0.5, 0.61, 0.7, 0.78, 0.85, 0.91, 0.96, 1.0\}$ and

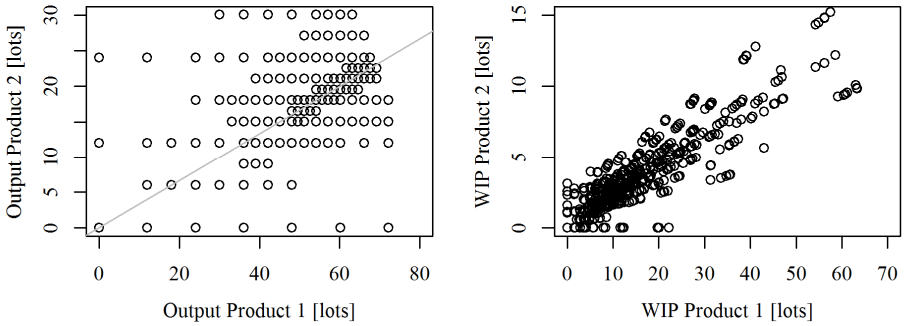

Figure 2: System state set of type Vargrid with base step size 12 (Vargrid12).

calculate corresponding average release quantities R_{gu} for products $g \in G$ and utilization levels $u \in U$. We generate samples

$$X_{gu} := R_{gu}(1 + r), g \in G, u \in U, \tag{19}$$

where r is a realization of the random variable $R \sim N(0, \sigma^2)$ with $\sigma = 0.25$. This process is repeated until the desired number of samples is reached. The resulting distributions of system states for 400 samples is shown in Figure 3.

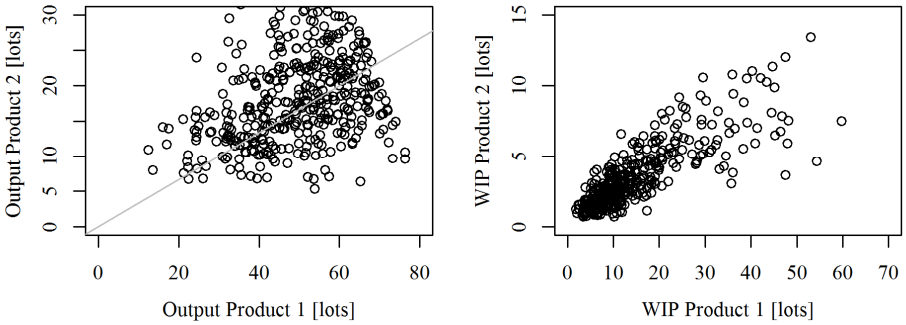

Figure 3: System state set of type stochastic with 400 samples (Stochastic400).

Table 2 provides an overview of all generated system state sets. Grid-based sets are larger to compensate for the higher coverage of areas with low utilization and high deviation from the product mix.

Table 2: Overview of all generated system state sets.

Size/Type	Grid	Vargrid	Stochastic
medium	Grid7 (280 states)	Vargrid16 (205 states)	Stochastic160 (160 states)
large	Grid5 (749 states)	Vargrid12 (520 states)	Stochastic400 (400 states)

5 COMPUTATIONAL EXPERIMENTS

5.1 Design of Experiments

The ACF model serves as a reference for achievable total cost and profit values. Its computing time is less than one second per instance for the Kayton model. We consider a planning horizon of $T = 15$ weeks. The optimization model is initialized with WIP levels corresponding to the recorded lots from long simulation runs at the prescribed BNU level. Three periods are added at the end of the planning horizon to account for end of horizon effects. Demand is generated for these periods by taking the averages of the previous three periods (Kacar et al. 2016). Moreover, backlog cost is multiplied by 5 for the last period. Costs are set to $\omega_{gt} = 35$, $b_{gt} = 50$, and $h_{gt} = 15$ while revenue is set to 20 for each completed lot. We examine four scenarios with normally distributed demand for a BNU of 70% and 90% with a coefficient of variation (CV) of 0.1 and 0.25 with five independent realizations each. The models are parameterized with the CF used by Gopalswamy and Uzsoy (2018) or a state set. We use $m = 1$ in (17). A computing time limit of 300 seconds per instance is applied. The lots to be released in a period are distributed uniformly over this period. We initialize the simulation model with the same initial WIP as previously used for the optimization model. Ten independent simulation replications are carried out for executing each release schedule to determine expected cost and revenue values. Table 3 summarizes the experimental design. The experiments are conducted on an Intel(R) Core(TM) i7-8700 CPU 3.20GHz PC with 16 GB RAM. IBM ILOG CPLEX 12.7.1 is used for solving the planning models. The simulation runs are carried out using AutoSched AP 11.3.0. The planning models and the infrastructure are coded in the C++ programming language.

Table 3: Design of experiments.

Factor	Level	Count
Planning formulation	(ACF,) DD	2
WIP Point	W_t, W_{t-1}, W_{avg}	3
Release quantity limit	None, Capacity (C)	2
System state set	Grid7, Grid5, Vargrid16, Vargrid12, Stochastic160, Stochastic400	6
Planned BNU	70%, 90%	2
CV of demand	0.1, 0.25	2
Demand realizations		5
Simulation replications		10
Total simulation runs		7400

5.2 Simulation Results

The performance of the DD variants combined with different state sets is examined relative to the ACF model. Table 4 summarizes the realized total cost and profit for all demand scenarios, each being averaged over five demand realizations. The total profit considers the realized revenue. To make the comparison easier, the sum of the cost and profit values over all demand scenarios are shown in the last two columns. The best 30% of the realized value ranges for cost and profit of each scenario are shaded in green, darker shades represent better results.

The ACF model outperforms the DD models for BNU=70%, CV=0.25 as well for BNU=90%, CV=0.1. At least one DD variant reaches a lower total cost and higher profit in the other two cases. Among all DD variants and system state sets we are not able to identify a variant that is clearly superior. Considering the WIP points, W_{avg} performs worse than W_{t-1} and W_t. The WIP levels prescribed by the chosen system states for all periods have to be met by the averages of the beginning and end of each period. This leads to an undulating progression with high fluctuations in release quantities. While W_t works well for low BNU and CV values, W_{t-1} performs better in all other scenarios. This is consistent with the expectation that the WIP point at the end of the period requires cycle times to be small enough compared to the period length to reach steady state within that period. Otherwise the WIP at the beginning of that period has a higher influence on the output. The release quantity limit reduces the problems with the W_{avg} setting and leads to slightly better results with W_t. We observe slightly increasing total cost for W_{t-1}. Larger system state sets enable a more exact adjustment of output values to the given demand patterns. Grid5, Vargrid12, and Stochastic400 perform better than their smaller counterparts, despite a higher computational burden and a time limit of 300 seconds. The overall best results are achieved with the Grid5 set, despite a lower density of states for high utilization around the product mix compared to the other set types. Analyzing the results in more detail, we notice that the performance of the DD variants and system state sets is closely related to how much the average WIP cost obtained by the simulation replications deviates from the optimization model values.

The actual WIP cost is only 1.98% higher than expected for DD_W_{t-1}_Grid5, whereas it is 5.32% and 6.06% higher for DD_W_{t-1}_Vargrid12 and DD_W_{t-1}_Stochastic400. While a higher density of system states might enable a better adjustment to demand, it might also lead to more variability and transient behavior with an increased WIP cost to output ratio. How to consider this in the model and mitigate its adverse effects on performance is part of future research. Differences in computing time and MIP gap as a result of factor level variations are depicted in Figure 4. The values are calculated by averaging over all DD variants and state sets for the respective factor levels. A positive gap implies that an optimality proof was impossible for at least one problem instance within the given time limit. The computing time is higher for the larger state set of each type. Solving instances with set type Stochastic takes longer than with Vargrid, which takes longer than those with Grid, despite an increasing set size.

Table 4: Results of the ACF model and the DD variants for all demand scenarios.

Model	State Set	WIP Point	X Limit	u70 cv10		u70 cv25		u90 cv10		u90 cv25		Sum	
				Cost	Profit	Cost	Profit	Cost	Profit	Cost	Profit	Cost	Profit
ACF				8930	15016	9499	14039	17575	13037	20685	9507	56689	51599
DD	Grid7	W_{avg}		10280	13642	10701	12943	21543	8955	26057	3882	68580	39422
			C	10292	13652	10402	13204	21029	9606	23300	6839	65023	43301
		W_{t-1}		9423	14566	9887	13741	18544	12142	20914	9411	58768	49861
			C	9414	14588	9895	13731	18815	11896	20917	9370	59041	49586
		W_t		9576	14394	9877	13752	19444	11163	22376	7876	61272	47185
			C	9548	14428	9907	13734	18332	12346	20236	9908	58023	50416
	Grid5	W_{avg}		9877	14014	10374	13221	23968	6632	24561	5445	68781	39312
			C	9716	14218	10359	13185	19236	11482	21204	9094	60516	47979
		W_{t-1}		9193	14777	9823	13726	17782	12989	20247	9953	57045	51445
			C	9048	14915	9865	13695	18319	12455	20221	9989	57454	51053
		W_t		9113	14833	9844	13676	19662	10961	21373	8776	59993	48246
			C	9072	14873	10002	13552	19228	11411	20885	9290	59188	49125
	Vargrid16	W_{avg}		9974	13912	10670	12943	21464	8942	23420	6610	65528	42406
			C	9761	14208	10409	13165	19248	11404	20911	9302	60329	48079
		W_{t-1}		9058	14886	9894	13629	18472	12113	20477	9740	57901	50368
			C	9091	14858	9883	13625	18349	12278	20504	9704	57828	50465
		W_t		8803	15139	9822	13684	18781	11787	20893	9219	58299	49829
			C	8792	15136	9867	13660	19088	11484	20602	9472	58348	49751
	Vargrid12	W_{avg}		10019	13815	10947	12572	21906	8521	23235	6881	66106	41789
			C	10022	13836	10489	12999	18688	11862	21479	8663	60677	47360
		W_{t-1}		8870	15055	9865	13638	18074	12516	20354	9795	57163	51004
			C	8860	15046	9818	13688	18066	12532	20655	9516	57400	50783
		W_t		8877	15041	10026	13475	19024	11513	21064	9044	58991	49073
			C	8847	15076	9893	13608	18731	11867	21248	8838	58719	49388
	Stochastic160	W_{avg}		9727	14177	10423	13204	21798	8752	23341	6698	65289	42831
			C	10367	13429	10212	13329	18758	11893	21456	8825	60793	47477
		W_{t-1}		9159	14792	9979	13547	18394	12235	20251	9918	57784	50491
			C	9162	14796	10158	13375	18062	12607	20423	9789	57805	50567
		W_t		8982	14951	10075	13422	19137	11532	21051	9094	59246	48999
			C	9075	14848	10015	13453	18875	11805	20447	9680	58412	49786
	Stochastic400	W_{avg}		10576	13272	10417	13101	22179	8246	22626	7428	65798	42047
			C	10481	13373	10228	13337	18485	12198	21880	8351	61074	47259
		W_{t-1}		8979	14955	10017	13456	18010	12606	20066	10078	57072	51095
			C	9039	14882	10056	13461	18210	12368	19951	10232	57255	50943
		W_t		8883	15018	9751	13758	19214	11367	21492	8647	59340	48791
			C	8797	15130	9774	13704	18339	12275	21437	8713	58347	49823

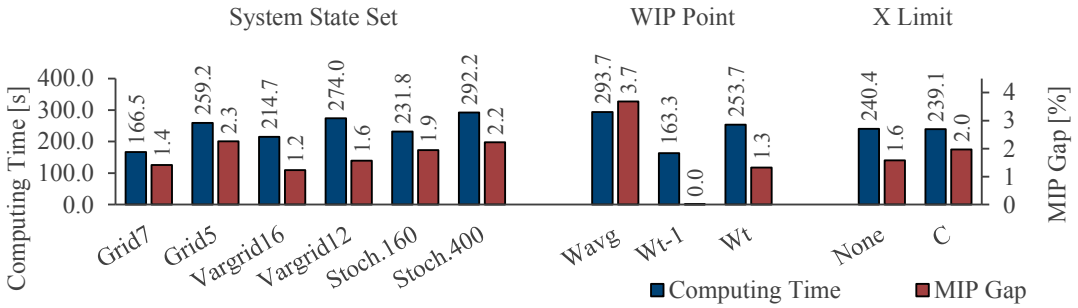

Figure 4: Factor level influence on computing time and MIP gap.

The largest MIP gap occurs for Grid5, despite Vargrid12 and Stochastic400 have higher computing times. W_{avg} takes longer to solve and has a larger MIP gap compared to the other WIP point settings. The dependency of WIP values in subsequent periods makes it harder to solve. Interestingly, all problem instances with W_{t-1} are solved to optimality whereas an average gap of 1.3% remains for W_t. The main difference in terms of computational tractability seems to be the different determination of the system state for the first period. It is fitted to the initial WIP in the W_{t-1} variants while for W_{t-1} it is chosen to minimize the total costs. The limit for release quantities does not change the computing time, but increases the average MIP gap slightly.

6 CONCLUSIONS AND FUTURE RESEARCH

We discussed DD formulations for wafer fabs. Several modifications of the basic DD formulation of Omar et al. (2017) were suggested. Moreover, different methods to determine system state sets were proposed. The planning models were assessed by executing the production plans using a simulation model of a scaled-down wafer fab. The simulation results demonstrated that variants of the DD model under certain experimental conditions are able to provide production plans having a very similar performance as the corresponding production plans obtained by the ACF model.

There are several directions for future research. It is desirable to repeat the experiments for a simulation model of a large-scaled wafer fab and for more general demand pattern including correlation between products and across periods. This includes experiments for multi-product settings. As a second research avenue, we are interested in modifying DD formulations in such a way that they can deal with the situation that the period length is smaller than the average cycle time. More research is also needed to determine for given demand minimal sets of state sets that lead to high-quality production plans. We believe that machine learning can be used for this task. A final direction is given by extending DD models towards integrated planning formulation for production and engineering activities (cf. Ziarnetzky and Mönch 2016).

ACKNOWLEDGMENTS

The authors would like to thank Karthick Gopalswamy, NCSU, for providing the CFs used in this research. The research was partially supported by the iDev 4.0 project. 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 programme. It is co-funded by the consortium members, grants from Austria, Germany, Belgium, Italy, Spain and Romania. The authors gratefully acknowledge the provided financial support.

REFERENCES

Asmundsson, J. M., R. L. Rardin, C. H. Turkseven, and R. Uzsoy 2009. "Production Planning Models with Resources Subject to Congestion". *Naval Research Logistics* 56:142-157.

Buzacott, J. A., and J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems*. Englewood Cliffs, NJ: Prentice-Hall.

Chien, C.-F., S. Dauzere-Peres, H. Ehm, J. W. Fowler, Z. Jiang, S. Krishnaswamy, L. Mönch, and R. Uzsoy. 2011. "Modeling and Analysis of Semiconductor Manufacturing in a Shrinking World: Challenges and Successes". *European Journal of Industrial Engineering* 5(3):254-271.

Fowler, J. W., L. Mönch, and T. Ponsignon. 2015. "Discrete-event Simulation for Semiconductor Wafer Fabrication Facilities: A Tutorial". *International Journal of Industrial Engineering: Theory, Applications, and Practice* 22(5):661-682.

Gopalswamy, K., and R. Uzsoy. 2018. "An Exploratory Comparison of Clearing Function and Data-driven Production Planning Models". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3482-3493. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gopalswamy, K., and R. Uzsoy. 2019. "A Data-driven Iterative Refinement Approach for Estimating Clearing Functions from Simulation Models of Production Systems". *International. Journal of Production Research* 57(19): 6013-6030.

Häussler, S., C. Stampfer, and H. Missbauer. 2020. "Comparison of Two Optimization based Order Release Models with Fixed and Variable Lead Times". *International Journal of Production Economics* 227:107682.

Kacar, N. B., D. F. Irdem, and R. Uzsoy. 2012. "An Experimental Comparison of Production Planning Using Clearing Functions and Iterative Linear Programming-Simulation Algorithms". *IEEE Transactions on Semiconductor Manufacturing* 25(1):104-117.

Kacar, N. B., L. Mönch, and R. Uzsoy. 2013. "Planning Wafer Starts Using Nonlinear Clearing Functions: A Large-Scale Experiment". *IEEE Transactions on Semiconductor Manufacturing* 26(4):602-612.

Kacar, N. B., L. Mönch, and R. Uzsoy. 2016. "Modeling Cycle Times in Production Planning Models for Wafer Fabrication". *IEEE Transactions on Semiconductor Manufacturing* 29(2):153 - 167.

Kacar, N. B., and R. Uzsoy. 2015. "Estimating Clearing Functions for Production Resources Using Simulation Optimization". *IEEE Transactions on Automation Science and Engineering* 12(2):539-552.

Kayton, D., T. Teyner, C. Schwartz, and R. Uzsoy. 1997. "Focusing Maintenance Improvement Efforts in a Wafer Fabrication Facility Operating Under the Theory of Constraints". *Production and Inventory Management Journal* 38(4):51-57.

Kayton Model. 2021. https://p2schedgen.fernuni-hagen.de/index.php?id=simulation&L=1. accessed 30th April 2021.

Missbauer, H. 2020. "Order Release Planning by Iterative Simulation and Linear Programming: Theoretical Foundation and Analysis of its Shortcomings". *European Journal of Operational Research* 280(2):495 – 507.

Missbauer, H., and Uzsoy, R. 2020. *Production Planning with Capacitated Resources and Congestion*. Springer: New York.

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: Springer.

Mönch, L., R. Uzsoy, and J. W. Fowler. 2018. "A Survey of Semiconductor Supply Chain Models Part III: Master Planning, Production Planning, and Demand Fulfillment". *International Journal of Production Research* 56(13):4524-4545.

Omar, R. S. M., U. Venkatadri, C. Diallo, and S. Mrishih. 2017. "A Data-Driven Approach to Multi-Product Production Network Planning". *International Journal of Production Research* 55(23):7110–7134.

Suri, R., and R. Hildebrant. 1984. "Modeling Flexible Manufacturing Systems Using Mean-Value Analysis". *Journal of Manufacturing Systems* 3(1):27-38.

Voß, S., and D. Woodruff. 2006. *Introduction to Computational Optimization Models for Production Planning in a Supply Chain.* 2nd ed., New York: Springer.

Wu, K. 2005. "An Examination of Variability and its Basic Properties for a Factory". *IEEE Transactions on Semiconductor Manufacturing* 18(1):214-221.

Ziarnetzky, T., N. Kacar, L. Mönch, and R. Uzsoy. 2015. "Simulation-based Performance Assessment of Production Planning Formulations for Semiconductor Wafer Fabrication". 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, 2884-2895. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ziarnetzky, T., and L. Mönch. 2016. "Incorporating Engineering Process Improvement Activities Into Production Planning Formulations Using a Large-scale Wafer Fab Model". *International Journal of Production Research* 54(21):6416-6435.

AUTHOR BIOGRAPHIES

TOBIAS VÖLKER is a teaching and research assistant and a master student at the Chair of Enterprise-wide Software Systems, University of Hagen. He received a bachelor degree in Information Systems from the University of Hagen, Germany. His research interests include production planning, discrete-event simulation, and data science. His email address is Tobias.Voelker@fernuni-hagen.de.

LARS MÖNCH is Professor in the Department of Mathematics and Computer Science at the University of Hagen, Germany. He received a master's degree in applied mathematics and a Ph.D. in the same subject from the University of Göttingen, Germany. His current research interests are in simulation-based production control of semiconductor wafer fabrication facilities, applied optimization and artificial intelligence applications in manufacturing, logistics, and service operations. His email address is lars.moench@fernuni-hagen.de.