Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

DATA-DRIVEN PRODUCTION PLANNING FORMULATIONS WITH INVENTORY CONSIDERATIONS

Tobias Völker
Lars Mönch

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

ABSTRACT

Data-driven (DD) production planning formulations for semiconductor wafer fabrication facilities (wafer fabs) are studied in this paper. These formulations are based on a set of system states representing the congestion behavior of the wafer fab with work in process and resulting output levels. We establish two DD formulations with inventory considerations. The first variant is a shortfall-based chance-constrained formulation that considers safety stocks at the finished goods inventory level. The second variant is a simple scenario-based stochastic program where the objective function reflects the expected inventory holding and backlog cost under uncertainty. The two variants are compared with the conventional DD formulation in a rolling horizon environment using a simulation model of a large-scaled wafer fab. The simulation experiments demonstrate that the stochastic program achieves the largest profit under all experimental conditions.

1 INTRODUCTION

Integrated circuits (ICs) are produced layer by layer on silicon wafers using hundreds of expensive machines in wafer fabs. The machines are organized in work centers. The moving entities in wafer fabs are lots consisting of up to 50 wafers. Different types of processes, i.e. batch and serial, are common in wafer fabs. A batch is a group of lots that are processed at the same time on a machine (Mönch et al. 2013). Sequence-dependent set-up times, auxiliary resources, and tight customer due dates for a large number of products can be observed in wafer fabs (Mönch et al. 2013). The routes of the most advanced products can contain up to 800 process steps. The same work center is visited up to 40 times by a single lot, i.e., reentrant process flows occur. Cycle time (CT), defined as the time span between material being released into the wafer fab and its emergence as finished product is of the order of twelve weeks in modern wafer fabs.

CTs have to be explicitly taken into account in production planning formulations of wafer fabs since they are long (Mönch et al. 2018). The CT increases nonlinearly with resource utilization as can be seen by queuing theory, experiments with discrete-event simulation, and industrial observations. The utilization, however, is determined by the release decisions made by production planning. Therefore, the CTs should be treated as endogenous to the production planning problem, i.e. workload-dependent lead times, estimates of the CTs, have to be taken into account in production planning formulations.

In the present paper, we continue our study of data-driven (DD) formulations proposed by Omar et al. (2017). DD production planning formulations are based on a set of system states representing the congestion behavior of a wafer fab with work in progress (WIP) and resulting output levels. They can be seen as an alternative to clearing function (CF)-based production planning formulations (Missbauer and Uzsoy 2020). We establish DD production planning formulations that explicitly consider inventory and backlog subject

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

to process and demand uncertainty. We demonstrate by comparing the new DD formulations with the conventional one in a rolling horizon environment that it is worth to consider safety stocks and inventory and backlog uncertainty in DD formulations.

The paper is organized as follows. In the next section, we will describe the problem setting and discuss related work. The different production planning formulations will be established in Section 3. Computational results of the planning formulations applied in a rolling horizon setting will be presented in Section 4. Conclusions and future research directions will be discussed in Section 5.

2 DISCUSSION OF RELATED WORK AND PROBLEM STATEMENT

Production planning formulation based on nonlinear CFs can be seen as a parameterized approach. However, DD approaches (Omar et al. 2017, Gopalswamy and Uzsoy 2018) make the parameterization efforts for CFs to some extent obsolete. DD formulations are a planning approach based on choosing system states. System states consider different products and their relations. They take an aggregated view on the resources and process steps of the wafer fab. DD formulations 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.

The present authors have studied DD formulations in a series of papers. Approaches to determine an appropriate set of system states and different WIP-output relations in DD formulations are investigated by Völker and Mönch (2021) and (2023). Moreover, DD approaches are changed by Völker and Mönch (2022) in such a way that they are able to deal with situations where the period length in the planning model is smaller than the average CT.

Production planning and inventory management is rarely discussed in an integrated manner. A multi-stage stochastic programming model of a simplified wafer fab that includes a stochastic model of demand evolution over time is proposed by Higle and Kempf (2010). A production planning formulation subject to stochastic demand based on the additive martingale model of forecast evolution (MMFE) (Heath and Jackson 1994) that considers inventory, backorder, and shortfall costs using chance constraints to represent target service levels is established by Albey et al. (2015). Similar formulations are considered by Aouam and Uzsoy (2012), (2015), and Ravindran et al. (2011). Ziarnetzky et al. (2018), (2020) extend the formulation by Albey et al. (2015) for the additive MMFE towards the multiplicative one since technology improvements lead to new technology migrations that result in non-stationary demand for wafer fabs. Exogenous, fixed lead times that are an integer multiple of the period length and workload-dependent lead times based on CFs are used. The different formulations are tested in a rolling horizon environment using wafer fab simulation models of different sizes. The planning formulations with inventory considerations outperform the remaining ones under many experimental conditions.

Inspired by the superior performance of planning formulations including safety stock, we are interested in modifying DD formulations by considering safety stock or the uncertainty of finished goods inventory (FGI) and backlog. These formulations have to be assessed in a rolling horizon environment to allow for a more realistic performance assessment.

3 PRODUCTION PLANNING FORMULATIONS

3.1 DD Models

For the production planning problem, we assume a finite planning horizon consisting of T discrete planning periods of equal length. Demand information for each product $g \in G$ and planning period t is available. In addition, the initial WIP, FGI, and backlog have to be considered in planning.

DD planning models determine release schedules to satisfy demands based on a set of discrete system states $r \in R$ that characterize the nonlinear behavior of the system under consideration. Each state describes the expected WIP and throughput (TH) levels for each product under steady-state conditions. By selecting a state for each planning period $t = 1, \ldots, T$, the expected output levels over the planning horizon are adjusted. Release quantities are set to achieve the associated WIP levels.

The accuracy of DD models depends on establishing an appropriate temporal relationship between WIP and expected TH of the system states. Völker and Mönch (2023) show that given sufficiently long CTs, especially in the case of $CT \geq 1$ periods, modeling output as a function of the WIP at the beginning of the period yields better results as using the WIP at the end of the period. The basic DD formulation requires the following notation:

Sets and indices

t: period index
g: product index
G: set of all products g
r: state index
R: set of all system states r

Decision variables

Y_{gt}: output of product g in period t from the last operation of its routing
X_{gt}: quantity of product g released in period t
W_{gt}: WIP of product g at the end of period t
I_{gt}: FGI of product g at the end of period t
B_{gt}: backlog of product g at the end of period t
Γ_{rt}: binary variable taking on the value 1, if system state r is selected in period t and 0 otherwise

Parameters

ω_{gt}: unit WIP cost for product g in period t
h_{gt}: unit FGI holding cost for product g in period t
b_{gt}: unit backlogging cost for product g in period t
D_{gt}: demand for product g in period t
Q_{gr}: WIP level of product g in system state r
O_{gr}: expected output quantities of product g in system state r.

The DD formulation is given as follows:

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

subject to

$$W_{g,t-1} + X_{gt} - Y_{gt} = W_{gt}, \qquad g \in G, t = 1, \ldots, T \tag{2}$$

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

$$\textstyle\sum_{r \in R} Q_{gr} \Gamma_{rt} \leq W_{g,t-1}, \qquad g \in G, t = 1, \ldots, T \tag{4}$$

$$\textstyle\sum_{r \in R} O_{gr} \Gamma_{rt} = Y_{gt}, \qquad g \in G, t = 1, \ldots, T \tag{5}$$

$$\textstyle\sum_{r \in R} \Gamma_{rt} = 1, \qquad t = 1, \ldots, T \tag{6}$$

$$\Gamma_{rt} \in \{0,1\}, \qquad r \in R, t = 1, \ldots, T \tag{7}$$

$$W_{gt}, I_{gt}, B_{gt}, X_{gt}, Y_{gt}, \geq 0, \qquad g \in G, t = 1, \ldots, T. \tag{8}$$

The objective function (1) is the sum of WIP, FGI, and backlog costs. The material balance equations (2) and (3) model the changes in WIP, FGI and backlog over the planning horizon. Constraints (4) and (5) determine the WIP at the beginning of each period t and the expected output during t as a result of the selected system state. Constraint set (4) is modeled as an inequality to avoid infeasibility at high WIP levels during the rolling horizon planning. Due to equations (6) and the binary restriction (7), exactly one system state is selected for each period. The remaining decision variables are nonnegative due to constraint set (8).

3.2 Model Extensions with Inventory Considerations

The basic DD formulation (1)-(8) assumes deterministic output values Y_{gt} and deterministic demand quantities D_{gt} for each product g and planning period t. However, if the computed release schedules are implemented in a simulation model or a real production system, the output is stochastic. Therefore, at the time of planning, output must be modeled as a random variable $\tilde{Y}_{gt}$. Similarly, since demand information is updated over time, demand forecasts must also be considered as a random variable $\widetilde{D}_{gt}$. Inventory and backlog at the end of period t can then be modeled based on the stochastic output and demand of previous periods. Using the material balance equations (2), we obtain:

$$\tilde{I}_{gt} - \tilde{B}_{gt} = I_{g0} + B_{g0} + \sum_{\tau=1}^{t} \tilde{Y}_{g\tau} - \sum_{\tau=1}^{t} \widetilde{D}_{g\tau},$$

where $\tilde{I}_{gt}-\tilde{B}_{gt}$ is a random variable. For a pair of realizations of both $\tilde{I}_{gt}$ and $\tilde{B}_{gt}$, only one of them can be positive at the same time. Subtracting the deterministic planning values yields:

$$\tilde{I}_{gt} - \tilde{B}_{gt} - (I_{gt} - B_{gt}) = \sum_{\tau=1}^{t}(\tilde{Y}_{g\tau} - Y_{gt}) - \sum_{\tau=1}^{t}(\widetilde{D}_{g\tau} - D_{gt}).$$

To determine the distribution of $\tilde{I}_{gt} - \tilde{B}_{gt} - (I_{gt} - B_{gt})$, we make the following assumptions:

1. The output predictions Y_{gt} and demand forecasts D_{gt} in the planning and demand forecast models are unbiased. As a result, we get $E[\tilde{Y}_{gt}] = Y_{gt}$, $E[\widetilde{D}_{gt}] = D_{gt}$, and consequently $E[\tilde{I}_{gt} - \tilde{B}_{gt}] = I_{gt} - B_{gt}$.
2. The sum of demand quantities $\sum_{\tau=1}^{t} \widetilde{D}_{g\tau}$ are normally distributed with a standard deviation of $\sigma_{gt}^{(D)}$ that depends on the demand forecast model.
3. The cumulative output quantities $\sum_{\tau=1}^{t} \tilde{Y}_{g\tau}$ are normally distributed as well. However, in this case, the distribution does not directly depend on t, but rather on the selected state r in period t with a standard deviation of $\sigma_{gr}^{(Y)}$.
4. The random variables $\sum_{\tau=1}^{t}(\tilde{Y}_{g\tau} - Y_{g\tau})$ and $\sum_{\tau=1}^{t}(\widetilde{D}_{g\tau} - D_{g\tau})$ are independent, i.e., with $Y_{g\tau}$ and $D_{g\tau}$ taking on constant values, we have $Cov\left(\sum_{\tau=1}^{t} \tilde{Y}_{g\tau}, \sum_{\tau=1}^{t} \widetilde{D}_{g\tau}\right) = 0$.

Given these assumptions, the distributions of the random variables that describe the deviation from the planned inventory and backlog quantities can be stated as:

$$\tilde{I}_{gt} - \tilde{B}_{gt} - (I_{gt} - B_{gt}) \sim N\left(0, {\sigma_{gtr}^{(I-B)}}^2\right) \tag{9}$$

with ${\sigma_{gtr}^{(I-B)}}^2 = {\sigma_{gr}^{(Y)}}^2 + {\sigma_{gt}^{(D)}}^2$ if state r is selected for period t in the planning model ($\Gamma_{rt} = 1$).

Failing to consider the distributions of inventories and backlog in planning will result in an incorrect cost function and an overestimation of the achievable service levels, e.g. the probability of no backlog in a period or the fraction of demand that is not backlogged. While the assumptions made for the distributions in (9) should be viewed critically, the added information could be sufficient to improve planning results, even if the assumptions are violated. Determining more accurate distributions would allow for further improvements. In the following, we propose two different extensions to the DD formulation to account for stochastic inventory and backlog values as a result of uncertainty in production and demand forecasting.

In the first variant, we determine a safety stock level S_{gt} with $I_{gt} - B_{gt} \geq S_{gt}$ such that the probability of a shortfall greater than S_{gt} is at most equal to a given value $1 - \vartheta$, i.e. we have $P\big(\tilde{I}_{gt} - \tilde{B}_{gt} - \big(I_{gt} - B_{gt}\big) \leq -S_{gt}\big) = 1 - \vartheta$ which is equivalent to $P\big(\tilde{B}_{gt} - \tilde{I}_{gt} - \big(B_{gt} - I_{gt}\big) \leq S_{gt}\big) = \vartheta$. The parameter ϑ can be interpreted as the target service level. We calculate the appropriate stock level using:

$$S_{gt} = F^{-1}_{\tilde{I}_{gt} - \tilde{B}_{gt} - (I_{gt} - B_{gt})}(\vartheta) = F^{-1}_{N(0,1)}(\vartheta)\, \sigma^{(I-B)}_{gtr}, \tag{10}$$

where $F^{-1}_{N(0,1)}(\cdot)$ is the quantile function of the standard normal distribution. Without a closed form representation, we use an approximation algorithm to determine its value (Acklam 2003). In a rolling horizon setting with unrestricted demand values and limited production capacities, we must allow for a planned shortfall U_{gt} to avoid infeasibility of the planning model, resulting in the chance constraint (CC) $B_{gt} - I_{gt} + S_{gt} \leq U_{gt}$. To avoid the occurrence of positive values for U_{gt} unless necessary, it must be penalized with an appropriate cost factor u_{gt} in the objective function (Albey et al. 2015). Note that the actual shortfall is a random variable $\widetilde{U}_{gt} = \max\big(0, \tilde{B}_{gt} - \tilde{I}_{gt} + S_{gt}\big)$.

The resulting planning model considers safety stocks in a similar way as the simple rounding down with safety stock (SRD-SS) and allocated clearing function with safety stock (ACF-SS) formulations in Ziarnetzky et al. (2020) with CT-driven safety stock settings. However, the value of S_{gt} is determined by the selected system state for period t. Consequently, the safety stock levels do not need to be a fixed fraction of the expected output based on the CT distribution. The DD-CC variant uses the following additional notation:

Decision variables

S_{gt}: target safety stock level for product g at the end of period t

U_{gt}: planned shortfall for product g at the end of period t

Parameters

ϑ: target service level

u_{gt} unit shortfall cost for product g in period t

$\hat{\sigma}^{(I-B)}_{grt}$: estimate for the standard deviation of the random variable $\tilde{I}_{gt}$-$\tilde{B}_{gt}$ for product g in period t given system state r.

The DD-CC formulation uses the objective function

$$\min \sum_{g \in G} \sum_{t=1}^{T} \big(\omega_{gt} W_{gt} + h_{gt} I_{gt} + b_{gt} B_{gt} + u_{gt} U_{gt}\big) \tag{11}$$

and additional or modified constraints

$$B_{gt} - I_{gt} + S_{gt} \leq U_{gt}, \qquad g \in G, t = 1, \dots, T \tag{12}$$

$$F^{-1}_{N(0,1)}(\vartheta)\ \textstyle\sum_{r \in R} \hat{\sigma}^{(I-B)}_{grt} \Gamma_{rt} = S_{gt}, \qquad g \in G, t = 1, \dots, T \tag{13}$$

$$W_{gt}, I_{gt}, B_{gt}, X_{gt}, Y_{gt}, S_{gt}, U_{gt} \geq 0, \qquad g \in G, t = 1, \dots, T. \tag{14}$$

The new objective function (11) includes the planned shortfall costs $u_{gt}U_{gt}$ for each product g and period t. Constraints (12) determine the planned shortfall with respect to the safety stock level. Appropriate values for safety stocks are determined in constraint set (13) based on the estimated standard deviation for the normally distributed deviation from planned inventory and backlog values given the selected system state r for period t and the target service level. The decision variables S_{gt} and U_{gt} are added to the nonnegative conditions in constraint set (14) which replaces (8).

While the service level ϑ in (13) can be set to achieve a cost-minimal balance between expected inventory and backlog values, the objective function (11) itself is still not reflective of the true expected costs under uncertainty. The inventory and backlog costs are calculated as

$$\sum_{g\in G}\sum_{t=1}^{T}\left(h_{gt}I_{gt}+b_{gt}B_{gt}\right)=\sum_{g\in G}\sum_{t=1}^{T}\left(h_{gt}\max\left(0,E\left[\tilde{I}_{gt}-\tilde{B}_{gt}\right]\right)+b_{gt}\ \max\left(0,E\left[\tilde{B}_{gt}-\tilde{I}_{gt}\right]\right)\right).$$

However, the expected values of the random variables $\tilde{I}_{gt}$ and $\tilde{B}_{gt}$, where $E\left[\tilde{I}_{gt}\right]$ and $E[\tilde{B}_{gt}]$ can take on a positive value at the same time, have to be considered separately:

$$\sum_{g\in G}\sum_{t=1}^{T}\left(h_{gt}\,E\left[\tilde{I}_{gt}\right]+b_{gt}\,E[\tilde{B}_{gt}]\right).$$

The expected values can be approximated based on representative samples of $\tilde{I}_{gt}$ and $\tilde{B}_{gt}$. We generate a set of probabilities $P=\left\{\frac{1}{2N},\frac{3}{2N},\dots,\frac{2N-1}{2N}\right\}$ with N equidistant probability samples and use inverse transform sampling (Law 2007) to derive scenarios $I_{gt}^{p}-B_{gt}^{p}$ with $I_{gt}^{p},B_{gt}^{p}\geq 0$ such that

$$P(\tilde{I}_{gt}-\tilde{B}_{gt}\leq I_{gt}^{p}-B_{gt}^{p})=p\in P.$$

Similar to the determination of safety stocks in (10), this can be implemented using an approximation algorithm for the quantile function of the standard normal distribution to determine the offset between the scenario variables I_{gt}^{p}, B_{gt}^{p} and the regular planning variables I_{gt}, B_{gt}:

$$I_{gt}^{p}-B_{gt}^{p}-(I_{gt}-B_{gt})=F_{\tilde{I}_{gt}-\tilde{B}_{gt}-(I_{gt}-B_{gt})}^{-1}(p)=F_{N(0,1)}^{-1}(p)\,\sigma_{gtr}^{(I-B)}.$$

Considering the expected FGI and backlog values as the averages of all scenarios in the objective function yields a simple stochastic program (SP), where the scenarios $I_{gt}^{p}-B_{gt}^{p}$ do not carry over to subsequent periods via the material balance equation (3). This distinguishes the SP from the two-stage stochastic programming (2SP) formulation of Aouam and Uzsoy (2015) where the scenarios are modeled in the form of independent demand realizations for all periods with a separate material balance equation for each scenario. The DD-SP variant uses the objective function

$$\min\sum_{g\in G}\sum_{t=1}^{T}\left(\omega_{gt}W_{gt}+\frac{1}{|P|}\sum_{p\in P}\left(h_{gt}I_{gt}^{(p)}+b_{gt}B_{gt}^{(p)}\right)\right)\tag{15}$$

and additional or modified constraints

$$I_{gt}^{(p)}-B_{gt}^{(p)}-\left(I_{gt}-B_{gt}\right)=F_{N(0,1)}^{-1}(p)\,\hat{\sigma}_{gt}^{(I-B)},\qquad g\in G,t=1,\dots,T,p\in P\tag{16}$$

$$\sum_{r\in R}\hat{\sigma}_{grt}^{(I-B)}\Gamma_{rt}=\hat{\sigma}_{gt}^{(I-B)},\qquad g\in G,t=1,\dots,T\tag{17}$$

$$W_{gt}, I_{gt}, B_{gt}, X_{gt}, Y_{gt}, I_{gt}^{p}, B_{gt}^{p} \geq 0, \qquad g \in G, t = 1, \dots, T. \qquad (18)$$

The objective function (15) is the sum of WIP and expected FGI and backlog costs. Constraints (16) offset the scenario variables from the regular planning variables for FGI and backlog. The standard deviation for the deviation from I_{gt} and B_{gt} is determined based on the active system state r in period t in constraint set (17). The nonnegativity constraints (18) replace (8).

The differences between the basic DD formulation (1)-(8), the DD-CC variant (2)-(7), (11)-(14) and the DD-SP variant (2)-(7), (15)-(18) can be described based on the cost function for FGI, backlog, and shortfall dependent on the target inventory level. Figure 1 shows the differences for different values of the standard deviation $\sigma_{grt}^{(I-B)}$.

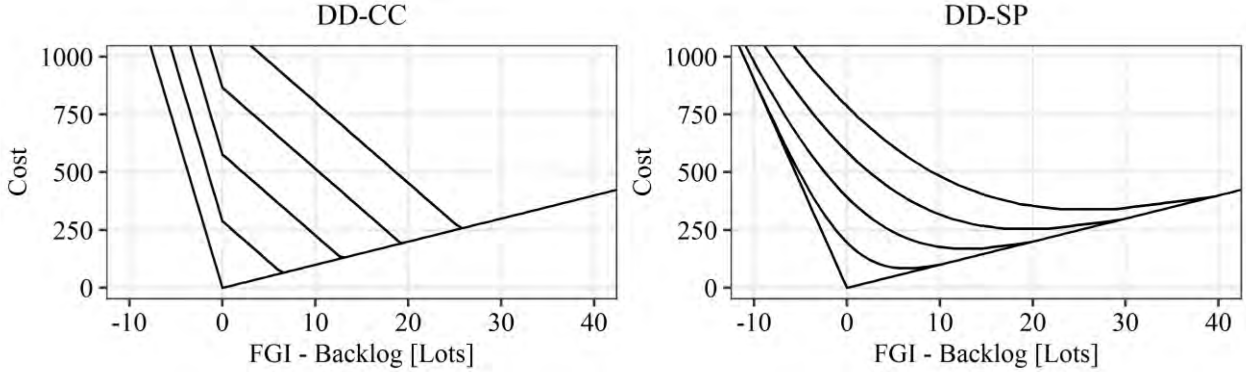

Figure 1: Total FGI, backlog and shortfall costs based on a target inventoy level $I_{gt} - B_{gt}$ in the DD-CC variant (left) and the DD-SP variant (right) using cost factors $h = 10$, $b = 90$, $u = 45$, standard deviation levels $\sigma^{(I-B)} = 0, 5, 10, 15, 20$, and a target service level of $\vartheta = 0.9$.

In the DD-CC variant, falling below the safety stock level results in a linear increase in costs. If the planned FGI is used up, additional costs are incurred for the backlog. The cost curves of the DD-SP variant are piecewise linear with $|P| + 1$ segments. They reach their minimum approximately at the safety stock levels of the DD-CC variant with a service level of $\vartheta = b/(b + h)$ where b and h represent backlog and FGI holding costs, respectively. Around these points, however, the costs initially remain at a similar level, allowing for greater planning flexibility.

4 SIMULATION BASED PERFORMANCE ASSESSMENT

4.1 Simulation Infrastructure and Simulation Model

The simulation experiments are conducted using the simulation infrastructure proposed by Ziarnetzky et al. (2015). A blackboard-type data layer forms the interface between the planning, control, and execution level. The execution level is represented by the simulation model. The data layer contains business objects such as machines and lots and is updated in an even-driven manner using notification functions of the simulation engine AutoSched AP 11.3 to reflect the current simulation state. The DD formulations are implemented as part of the planning level in the C++ programming language using the commercial solver IBM ILOG CPLEX 12.7.1. To compute a new release schedule, the control level extracts relevant information from the data layer and instantiates a new instance of the planning model. After the model is solved, the plan is executed by the control level which releases the specified number of lots uniformly over the respective periods. For the experiments, we use a rolling horizon planning approach. At the beginning of each planning epoch, a new release schedule is created. Only the first period of the plan is executed before replanning takes place.

The MIMAC I simulation model (Fowler and Robinson 1995) used in the experiments represents a large-scale wafer fab with more than 200 machines organized in 69 work centers. The steppers of the lithography area serve as a planned bottleneck. Processing characteristics include batch processing machines, sequence-dependent setup times, and operators. Exponentially distributed machine breakdowns are the major contributor to variability. We differentiate between scenarios with short and long machine failure durations. For long durations, the mean time to repair (MTTR) and the mean time to failure (MTTF) are set to be twice as long as in the short duration case. Two products are considered, each requiring over 200 process steps with highly reentrant process flows. First-In-First-Out (FIFO) dispatching is used. The processing times are deterministic.

4.2 Demand Generation

Demand is generated for two different demand types, each with planned bottleneck utilization (BNU) levels of 70% or 90% and a product mix (PM) of 1:1. First, using simulation, mean demand values are determined for the target BNU levels and the given product mix. These demand values are then modified depending on the demand type and BNU. For demand of type level load (ll), one of the products is selected with equal probability every three periods. For each of the next three periods, the mean demand for that product is increased by 5% at a target BNU of 90% or by 10% at a target BNU of 70%. Demand for the other product is reduced by 5% or 10%, respectively. As a result, the demand values for the two products are negatively correlated. For the time-varying load (tv) demand type, the mean demand for both products is randomly increased or decreased simultaneously for sets of three consecutive periods in a similar manner, resulting in positively correlated demand. Finally, normally distributed demand realizations around the modified mean demand values M_{gs} for every product g and planning epoch s are generated with a coefficient of variation of $CV = 0.25$:

$$D_{gs} \coloneqq M_{gt}\left(1 + r_{gs}\right),\ g \in G, s = 1, \ldots, H + T - 1,$$

where r_{gs} is a realization of the normally distributed random variable $R \sim N(0, \sigma^2)$ with $\sigma = CV$ and H is the length of the simulation horizon in periods. The demand values are known in advance. Accordingly, the experiments do not consider uncertain demand forecasts and instead focus on production uncertainty.

4.3 Determining System States

System states describe the relationship between the expected WIP and TH values under steady-state conditions. We use long-term simulation runs with constant release rates equal to the desired TH levels to derive corresponding WIP values for all products. The TH values must be sampled from the set of feasible values given the systems production capacity. We follow the HC-ipt sampling procedure proposed in Völker and Mönch (2023). A summary of the procedure is given below.

We start by creating a space-filling hypercube design using the intersite-proj-th method of Crombecq et al. (2011) with one sample $z_r \in [0,1]^{|G|}$ for each state $r \in R$. The first component $z_r^{(1)}$ of the sample is used to determine the BNU and consequently mean release quantities μ_r. The PM of μ_r is perturbed using a BNU-invariant vector Δ_r based on the remaining components $z_r^{(2)}, \ldots, z_r^{(|G|)}$. The system state sample, consisting of the release quantities $X_r \in \mathbb{R}^{|G|}$, is then given by $X_r = \mu_r + \Delta_r$.

We sample the BNU based on a triangular distribution with a lower limit of 0% and an upper limit and mode of 100%. The PM perturbation is normally distributed with a $CV = 0.25$, equal to the setting for the demand distribution. For both short and long machine breakdowns, $|R| = 200$ system states are generated. The samples are simulated over ten years after one year of warmup time to reach steady state. Recorded statistics include the WIP Q_{gr}, the TH O_{gr}, the CTs CT_{gr} and the standard deviations of WIP $\sigma_{gr}^{(W)}$ and CT $\sigma_{gr}^{(CT)}$. Figure 2 (left) shows the distribution of the sampled states in terms of TH. The ratio of Q_{gr}/O_{gr} can be interpreted as implied lead times (LTs), estimates for the expected CTs in each state. The LTs for short and long machine breakdowns are plotted as a function of the expected BNU in the center and on the right of Figure 2, respectively.

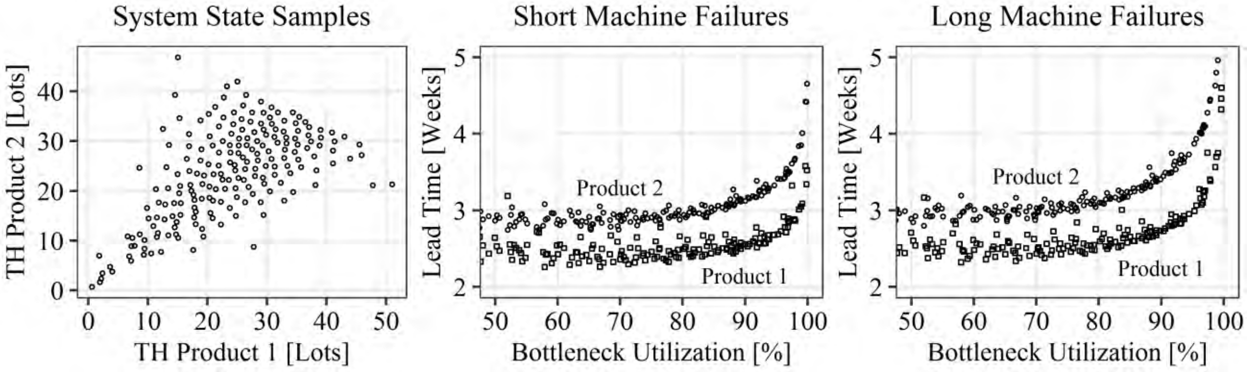

Figure 2: Sampling distribution of system states by expected TH (left) and implied lead times derived from simulated WIP values relative to the bottleneck utilization for short (center) and long (right) machine failure durations.

We estimate output uncertainty represented by $\sigma_{gr}^{(Y)}$ based on the data gathered for the system states. During each simulation run, the input rate X_{gt} remains constant, while the WIP values W_{gt} and the output quantities Y_{gt} vary over time. As long as the system is stable, the expected output quantities per period are equal to the input rate, i.e. $E[Y_{gt}] = E[Y_g] = X_{gt}$. We assume that the initial WIP differs from the expected WIP by κ, i.e., $W_{g0} = E[W_{gt}] + \kappa = E[W_g] + \kappa$ holds. Based on the WIP balance equation (2), we obtain $W_{gt} - W_{g0} = \sum_{\tau=1}^{t}(X_{g\tau} - Y_{g\tau})$ from which $W_{gt} - E[W_g] - \kappa = t \cdot E[Y_g] - \sum_{\tau=1}^{t} Y_{g\tau}$ follows. The deviation of the WIP from its expected value $W_{gt} - E[W_{gt}]$ is therefore equal to the negative deviation of the accumulated output quantities from their expected values, offset by the constant κ. Accordingly, the standard deviation of the WIP $\sigma_g^{(W)}$ can be used as a proxy for the standard deviation of the accumulated output quantities. We set $\hat{\sigma}_{gr}^{(Y)} \coloneqq \sigma_{gr}^{(W)}$. Table 1 provides an overview of the average standard deviation $\sigma_g^{(W)}$ for all system states within given utilization level intervals with short and long machine failure scenarios. Alternative measures of output uncertainty $\sigma_g^{(CT)} Y_g$ are provided for comparison.

Table 1: Measures for output uncertainty averaged over all system states in a bottleneck utilization interval.

Utilization Interval	Short Machine Failures				Long Machine Failures			
	$\sigma_1^{(W)}$	$\sigma_2^{(W)}$	$\sigma_1^{(CT)} Y_1$	$\sigma_2^{(CT)} Y_2$	$\sigma_1^{(W)}$	$\sigma_2^{(W)}$	$\sigma_1^{(CT)} Y_1$	$\sigma_2^{(CT)} Y_2$
(50%, 60%]	3.08	3.58	5.29	6.00	4.01	5.48	6.21	7.71
(60%, 70%]	3.46	4.19	5.76	6.73	4.86	7.10	7.15	9.36
(70%, 80%]	3.95	4.96	6.30	7.47	6.12	8.89	8.48	11.12
(80%, 90%]	4.61	6.03	7.00	8.53	7.53	11.37	9.90	13.59
(90%, 100%]	5.91	7.54	8.28	9.89	11.30	15.42	13.50	17.40

4.4 Design of Experiments

The simulation experiments are designed to determine whether the DD-CC and DD-SP extensions to the basic DD formulation can contribute to lower costs and higher profits under production uncertainty and certain demand forecasts. The production-side uncertainty is determined in particular by the length of machine failures and the expected bottleneck utilization. Accordingly, we conduct experiments with short and long failure durations and low (70%) and high (90%) levels of BNU. Level load and time varying load demand types are intended to show the impact of changes in product mix and utilization over time on the planning performance. For each of the resulting eight scenarios, ten demand realizations are generated and replicated in five independent simulation runs. The design of experiments is summarized in Table 2.

Table 2: Design of Experiments.

Factor	Level	Count
Planning models	DD, DD-CC, DD-SP	3
Demand type	ll, tv	2
Planned bottleneck utilization	70%, 90%	2
Machine failure duration	short, long	2
Demand realizations		10
Simulation replications		5
Total simulation runs		1200

The performance of the planning models is evaluated for simulation runs over 52 weeklong periods. Each simulation is initialized with a WIP snapshot, taken after one year of initial simulation time, that is unique to the respective scenario, demand realization and simulation replication. At the beginning of each planning epoch, the planning model is instantiated to calculate a new release schedule. The time to solve the models is limited to ten seconds, after which the best solution found will be returned. Each planning instance uses deterministic information on demand for twelve periods. Three additional periods with averaged demands serve to avoid end of horizon effects. The unit costs for WIP, FGI, backlog, and shortfall are set to $\omega_{gt} = 60$, $h_{gt} = 10$, $b_{gt} = 90$, and $u_{gt} = 45$, respectively. A unit revenue of 450 is used to calculate overall profits. We use a target service level of $\vartheta = b_{gt}/(h_{gt} + b_{gt}) = 0.9$ for the DD-CC variant. For DD-SP, $N = 20$ scenarios for FGI and backlog are used. The experiments are executed on an Intel® Core™ i7-8700 CPU 3.20GHz PC with 16GB RAM.

4.5 Computational Results

The results of the experiments are presented in Table 3. By taking into account stochastic FGIs and backlogs, both the DD-CC and the DD-SP variants can substantially reduce the sum of the corresponding costs by 13% to 30%, depending on the scenario, with higher levels of reduction at lower BNU. The reason for the reduction lies in the shift towards higher FGI levels, which reduces the occurrence of backlog at a much higher cost. The ratio of FGI to backlog costs is very similar for both variants. This can be explained by the fact that, despite the different objective functions and constraints, the resulting cost functions are also similar, as can be seen in Figure 1, since they reach their minimum at approximately the same inventory level. Although DD-SP tends to accumulate slightly more backlog than DD-CC, the sum of FGI and backlog costs are the lowest in six out of eight scenarios.

Smaller relative differences are found for WIP costs. However, since WIP is the dominant cost factor, the absolute differences are relevant. WIP costs for the DD-CC variant are generally higher than for the DD model. The requirement for higher safety stocks at higher utilization levels can cause larger fluctuations in the planned WIP, resulting in an increase of the average WIP due to the nonlinear relationship between WIP and TH. For the DD-SP variant, WIP costs are generally lower than for DD and DD-CC. Due to the shape of the cost function shown in Figure 1 (right), inventory levels can be varied around the cost minimum with only small differences in expected costs. As a result, capacity can be utilized more efficiently with respect to the expected WIP cost. Inventories are built in periods with otherwise low utilization and lowered in periods with high demand at little additional cost relative to the cost-optimal inventory level.

Overall, DD-CC achieves an increase in profit in all scenarios. However, the costs saved for FGI and backlog are partially offset by the higher WIP costs. DD-SP reduces FGI and backlog costs as well as the WIP costs, resulting in correspondingly higher profits. The increase in profits seems to be closely related to the degree of production-side uncertainty, with the largest differences being observed for long machine downtimes and a high BNU.

Table 3: Results of computational experiments.

Machine	Demand	Bottleneck	Planning	Costs			Profit	
Failures	Type	Utilization	Model	WIP	FGI	Backlog	Total	Change
short	ll	70%	DD	420,760.8	**5,244.8**	19,606.6	722,551.7	0.00%
			DD-CC	422,938.8	8,334.8	**9,607.5**	731,197.9	1.20%
			DD-SP	**418,880.4**	7,494.6	9,976.2	**734,494.8**	**1.65%**
		90%	DD	594,201.6	**7,307.3**	35,091.7	865,724.3	0.00%
			DD-CC	601,188.0	10,537.1	**26,082.3**	869,971.6	0.49%
			DD-SP	**593,205.6**	9,198.4	27,561.4	**875,932.5**	**1.18%**
	tv	70%	DD	421,412.4	**5,786.8**	19,462.1	713,825.6	0.00%
			DD-CC	422,178.0	9,012.8	**9,565.8**	724,482.4	1.49%
			DD-SP	**419,494.8**	7,940.5	10,189.2	**726,894.5**	**1.83%**
		90%	DD	595,183.2	**7,232.1**	35,054.2	860,796.4	0.00%
			DD-CC	601,666.8	10,700.5	**23,462.1**	868,286.6	0.87%
			DD-SP	**593,304.0**	9,203.6	24,780.5	**874,163.9**	**1.55%**
long	ll	70%	DD	441,421.2	**6,090.4**	26,196.0	693,655.4	0.00%
			DD-CC	443,898.0	11,429.2	**11,589.3**	708,204.5	2.10%
			DD-SP	**439,465.2**	10,202.6	12,471.0	**711,794.3**	**2.61%**
		90%	DD	667,563.6	**7,218.6**	76,748.1	747,086.6	0.00%
			DD-CC	677,944.8	12,071.2	60,140.9	756,083.0	1.20%
			DD-SP	**660,180.0**	9,797.7	**55,946.3**	**777,058.0**	**4.01%**
	tv	70%	DD	441,748.8	**6,928.1**	25,541.8	688,275.3	0.00%
			DD-CC	443,353.2	12,209.6	**11,725.9**	702,108.4	2.01%
			DD-SP	**439,532.4**	10,683.9	13,529.9	**704,291.8**	**2.33%**
		90%	DD	664,387.2	**7,409.8**	88,188.6	730,531.4	0.00%
			DD-CC	680,101.2	11,621.4	**63,943.0**	743,887.4	1.83%
			DD-SP	**658,615.2**	9,967.5	63,997.6	**764,164.7**	**4.60%**

5 CONCLUSION AND FUTURE RESEARCH

In this paper, we studied the performance of DD production planning models with inventory considerations. The proposed models were assessed in a rolling horizon environment that was provided by a simulation model of a large-scaled wafer fab. They were only confronted with production uncertainty which arises from machine breakdowns. The simulation experiments demonstrated that explicit inventory considerations in DD planning models lead to higher profits under all experimental conditions explored in the paper.

There are several directions for future research. First of all, we are interested in relaxing the assumptions made for the random variable that represents the difference of inventory and backlog. For instance, we are interested in applying the approaches to demand that follow the additive or multiplicative MMFE. It seems also interesting to extend the proposed models towards the generalized DD planning formulation proposed by Völker and Mönch (2022). Finally, it seems worthwhile to apply more advanced stochastic programming techniques to the DD planning formulations.

REFERENCES

Acklam, P. J. 2003. "An Algorithm for Computing the Inverse Normal Cumulative Distribution Function". https://web.archive.org/web/20151030215612/http://home.online.no/~pjacklam/notes/invnorm/. Last accessed 04.05.2023.

Albey, E., A. Norouzi, K. G. Kempf, and R. Uzsoy. 2015. "Demand Modeling with Forecast Evolution: an Application to Production Planning". *IEEE Transactions on Semiconductor Manufacturing* 28(3):374-384.

Aouam, T., and R. Uzsoy. 2012. "Chance-Constraint-Based Heuristics for Production Planning in the Face of Stochastic Demand and Workload-Dependent Lead Times". In *Decision Policies for Production Networks*, edited by K. Kempf and D. Armbruster, 173-208. Boston: Springer.

Aouam, T., and R. Uzsoy. 2015. "Zero-Order Production Planning Models with Stochastic Demand and Workload-Dependent Lead Times". *International Journal of Production Research* 53(6):1661-1679.

Crombecq, K., E. Laermans, and T. Dhaene. 2011. "Efficient Space-filling and Non-collapsing Sequential Design Strategies for Simulation-based Modeling". *European Journal of Operational Research* 214(3): 683-696.

Fowler, J. W., and J. Robinson. 1995. "Measurement and Improvement of Manufacturing Capacity (MIMAC) Final Report". Technology Transfer #95062861A-TR, SEMATECH.

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.

Heath, D., and P. Jackson. 1994. "Modeling the Evolution of Demand Forecasts with Application to Safety Stock Analysis in Production/Distribution Systems". *IIE Transactions* 26:17-30.

Higle, J., and K. Kempf. 2010. "Production Planning under Supply and Demand Uncertainty: A Stochastic Programming Approach". In *Stochastic Programming: The State of the Art*, edited by G. Infanger, 297-315. Berlin: Springer.

Law, A. M. 2007. *Simulation Modeling and Analysis.* 4th ed., Boston: McGraw-Hill.

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

Ravindran, A., K. Kempf, and R. Uzsoy. 2011. "Production Planning with Load-Dependent Lead Times and Safety Stocks for a Single Product". *International Journal of Planning and Scheduling* 1(1/2):58-86.

Völker, T, and L. Mönch. 2023. "Data-driven Production Planning Models for Wafer Fabs: An Exploratory Study". *IEEE Transactions on Semiconductor Manufacturing,* accepted.

Völker, T., and L. Mönch. 2021. "Data-driven Production Planning Formulations for Wafer Fabs: a Computational Study". In *Proceedings of the 2021 Winter Simulation Conference*, edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, 1-12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Völker, T., L. Mönch. 2022. "A Generalized Data-driven Production Planning Model: Algorithmic Foundation and Simulation-based Performance Assessment". In *Proceedings 22nd International Working Seminar on Production Economics*, Innsbruck.

Ziarnetzky, T., N. B. 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, V. W.K. Chan, I.-C. 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., L. Mönch, and R. Uzsoy. 2020. "Simulation-Based Performance Assessment of Production Planning Models with Safety Stock and Forecast Evolution in Semiconductor Wafer Fabrication". *IEEE Transactions on Semiconductor Manufacturing* 33(1):1-12.

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 in manufacturing. He can be reached by email at tobias.voelker@fernuni-hagen.de.

LARS MÖNCH is full professor of Computer Science at the Department of Mathematics and Computer Science, University of Hagen where he heads the Chair of Enterprise-wide Software Systems. He holds M.S. and Ph.D. degrees in Mathematics from the University of Göttingen, Germany. After his Ph.D., he obtained a habilitation degree in Information Systems from Technical University of Ilmenau, Germany. His research and teaching interests are in information systems for production and logistics, simulation, scheduling, and production planning. He can be reached by email at lars.moench@fernuni-hagen.de.