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

IMPACT OF PRODUCTION PLANNING APPROACHES ON WAFER FAB PERFORMANCE DURING PRODUCT MIX CHANGES

Tobias Völker
Lars Mönch

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

Reha Uzsoy

Edward P. Fitts Department of Industrial and Systems Engineering
North Carolina State University
4107 Fitts-Woolard Hall
Raleigh, NC 27695-7906, USA

ABSTRACT

We present the results of a series of simulation experiments examining the impact of product mix changes on global performance measures such as costs and profit. In these experiments, we apply three production planning models in a rolling horizon setting that differ in their anticipation of shop-floor behavior. The first two are based on exogenous, i.e. fixed, workload-independent lead times, while the third uses non-linear clearing functions to represent workload-dependent lead times. The simulation results clearly demonstrate the benefit of production planning models that correctly anticipate the queueing behavior of the wafer fab.

1 INTRODUCTION

Semiconductor wafer fabrication facilities (wafer fabs) are complex manufacturing systems that contain hundreds of complicated and expensive machines. In modern wafer fabs, products have process flows involving up to 500 different process steps. Several dozen different technologies exist in most wafer fabs. The product mix often changes over time as a result of incoming orders, different demand pattern, and the introduction and retirement (ramp-up and ramp-down) of new and old technologies. The product mix has a large impact on global fab performance measures such as cycle time (CT), throughput (TH) and on-time delivery performance (Mönch et al. 2013). The CT is the delay between work being released into a wafer fab and it emerging as output, and is of the order of ten weeks in most wafer fabs.

Product mix changes can result in non-stationary behavior of a wafer fab. Although production planning and shop-floor dispatching are important functions in semiconductor supply chains (Mönch et al. 2018a), their interaction is in general not well understood (Mönch et al. 2018b). The non-stationary behavior of a wafer fab during product mix changes is studied by Dümmler (2000), who analyzes the impact of short-term increases in wafer start rates (surges) and finds that an appropriate dispatching rule is important to ensure a fast recovery of the wafer fab. The effect of a frequently changing product mix on fab performance is investigated. However, production planning is considered only in a very basic way, ignoring limited capacity. In the present paper, we combine the investigations of Dümmler (2000) with different production planning approaches that anticipate shop-floor behavior in different ways. The first two models are based on exogenous, i.e. fixed, lead times that are an integer multiple of the period length or fractional, while the third uses non-linear clearing functions to represent workload-dependent lead times. Lead times (LT) are CT estimates that are applied in production planning. We are interested in demonstrating the benefit of better anticipation of shop-floor behavior.

The rest of the paper is organized as follows. The problem is described and analyzed in the next section, with a discussion of related work. The production planning approaches and LT estimates compared in this

paper are described in Section 3, and the simulation environment for the experiments in Section 4. The simulation experiments are presented in Section 5, and conclusions and future directions in Section 6.

2 PROBLEM DESCRIPTION AND ANALYSIS

2.1 Problem Statement

We investigate the short-term impact of product mix changes on global performance measures such as profit and costs in a single wafer fab. The wafer fab consists of machine groups where the machines belonging to a machine group provide the same functionality. We refer to these machine groups as work centers, and assume that we know the planned bottleneck work center of the wafer fab. Weekly periods are considered. Weekly demand, in number of wafers, is known for each product and period. This regular demand is chosen in such a way that a target bottleneck utilization is achieved. In order to model product mix changes, we increase the demand for one product for a prescribed number of periods. The demand signal that leads to the product mix change is called a surge impulse, and is characterized by its size, expressed as a multiple of the regular period demand, and a duration given in periods. We compare the following four production planning approaches with the surge impulse:

1. The simple backward planning approach used by Dümmler (2000).
2. A production planning model based on exogenous LTs $L(g)$ for product g that are an integer multiple of the period length. Since the integer lead times are estimated by rounding down the fractional lead time estimates obtained by simulation, we refer to this as the Simple Rounding Down (SRD) model (Kacar et al. 2012, Kacar et al. 2013).
3. The third planning model extends the previous one by taking into account fractional LTs and is therefore referred to as FLT model (Kacar et al. 2016).
4. The fourth model used is the Allocated Clearing Function (ACF) formulation of Asmundsson et al. (2006) and Asmundsson et al. (2009). Clearing functions (CFs) relate the expected output of a work center in a planning period to some measure of its expected workload over that period (Missbauer and Uzsoy 2011). The ACF model is based on the idea that the output of the different work centers of the wafer fab is estimated using an aggregate workload measure, and is then allocated to individual products.

Note that the four planning models differ in how they anticipate the behavior of the shop-floor. The backward planning model considers the finite capacity of the wafer fab only in a very simplistic way. Although the SRD and FLT models consider the finite capacity, they have the limitation that the LT does not depend on the workload. The ACF model respects the finite capacity of the shop-floor and is able to deal with the congestion of the shop-floor since CFs are used to represent the behavior of the individual work centers. Since it is unlikely that the wafer fab is in steady state after the surge impulse, we have to perform a transient analysis, i.e., we have to gather the performance measure values over time.

2.2 Previous Related Work

The impact of wafer surges is studied by Nag and Maly (1995) using a simulation model of a wafer fab. Rose (1998) uses discrete-event simulation to study the transient behavior of a wafer fab after a breakdown of the bottleneck work center. Only the bottleneck work center is modeled in detail. Dümmler (2000) and Dümmler and Rose (2000) use simulation to study the non-stationary behavior of a wafer fab after product mix changes. Only an extremely crude production planning scheme is used in the simulation experiments.

The second stream of research deals with planning models that are confronted with demand peaks, such as those from seasonally demanded goods. Several strategies for production planning with seasonal demand are proposed by Metter (1997) and Metter (1998). One obvious strategy is to build inventory during low level seasons with the goal to absorb demand peaks. Gangsterer (2015) demonstrates by simulation that a

demand pattern with large peaks is the most challenging demand scenario with respect to planning robustness, and that demand peaks can result in low service levels.

Production planning models for wafer fabs are proposed, for instance, by Leachman (2001), Kacar et al. (2013), and Kacar et al. (2016). They are applied in a rolling horizon setting among others by Ziarnetzky et al. (2018) and Ziarnetzky et al. (2020). Although quite general demand patterns based on the Martingale Model of Forecast Evolution (MMFE) (Heath and Jackson 1994, Norouzi and Uzsoy 2014) are used in the rolling horizon experiments, to the best of our knowledge, demand settings with large-sized peaks have not been considered until now. In the present paper, we extend the setting in Dümmler (2000) by confronting the SRD, FLT, and ACF planning models in a rolling horizon setting with a surge impulse in the demand.

3 PRODUCTION PLANNING AND CONTROL APPROACHES

3.1 Production Planning Models

We consider a simple backwards planning approach and three optimization-based planning formulations to determine release schedules based on given demand forecasts D_{gt} for every product g and period t. The backwards termination approach uses a constant LT estimate $L(g)$ for each product g to release material in the amount of $D_{g,t+\lfloor L(g) \rfloor}$ in each period t. It does not consider the current work in process (WIP), finished goods inventory (FGI), backlog or updates to demand forecasts and is therefore not expected to yield competitive results in terms of cost on the planning level. It serves as a reference (REF) approach, comparable to the fixed input rates in the experiments of Dümmler (2000).

The optimization-based planning models differ in their representation of the LTs. The SRD model uses integer LTs that establish a direct relationship between specific input and output periods, derived by rounding down given fractional LTs. The FLT model represents LTs more accurately, but requires additional constraints to model changes in output rates within a period. The ACF model includes endogenous LTs implicitly represented by the CFs. The resulting LTs depend on the initial state of the system and the evolution of estimated workload at each work center over time. A finite planning window of length T divided into discrete periods of equal length is considered. The objective of the models is to determine release quantities for each product and period which minimize costs; extensive discussions of the different models are given in Missbauer and Uzsoy (2020). The following notation is used for the SRD model:

Sets and indices

- G: set of all products
- K: set of all work centers
- t: period index
- g: product index
- k: work center index
- l: operation index
- $O(g)$: set of all operations of product g
- $O(g,k)$: set of all operations of product g performed on machines of 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_{gt}: quantity of product g released into the first work center in its routing 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

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 during period t
C_k: capacity of work center k in units of time
α_{gl}: processing time for operation l of product g
$L(g,l)$: time span from the release of the material to the completion of operation l of product g.

The model itself can be stated as follows:

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

subject to

$$W_{g,t-1} + X_{gt} - Y_{gt} = W_{gt}, \qquad g \in G, t = 1,\dots,T \quad (2)$$

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

$$Y_{gtl} = X_{g,t-\lfloor L(g,l)\rfloor}, \qquad g \in G, t = 1,\dots,T, l \in O(g) \quad (4)$$

$$\sum_{g\in G}\sum_{l\in O(g,k)} \alpha_{gl}Y_{gtl} \le C_k, \qquad k \in K, t = 1,\dots,T \quad (5)$$

$$X_{gt}, Y_{gtl}, Y_{gt}, W_{gt}, I_{gt}, B_{gt} \ge 0, \qquad g \in G, t = 1,\dots,T, l \in O(g). \quad (6)$$

The objective function (1) is the sum of WIP, inventory, and backlog cost. The constraint sets (2) and (3) enforce WIP and finished goods inventory balance. Constraints (4) are the input-output relations, (5) limit the capacity of the work centers, and (6) define the decision variables as nonnegative. The SRD model incorporates LT estimates $L(g,l)$ for each operation l of product g that are determined in a recursive manner, following Kacar et al. (2013). It requires the separate consideration of the capacity consumption and completion of the initial WIP $W_{g,0,l}$ in queue before operation l of product g at the beginning of the first period. $W_{g,0,l}$ represents material released in period $t = -\lfloor L(g,l-1)\rfloor$ and will therefore contribute to the output at operation $m \in O(g)$ in period $t = \lfloor L(g,m)\rfloor - \lfloor L(g,l-1)\rfloor$. Note that $W_{g,0,l}$ with $L(g,l-1) > \lfloor L(g)\rfloor$ will be released at the beginning of the first period without any additional capacity consumption. To avoid violating constraints (5), capacity usage by the initial WIP is modeled as a reduction of C_k down to a minimum of zero (Leachman 2001).

The FLT model allows Y_{gtl} to consist of material released in up to two periods, with the appropriate ratio determined by the fractional portion of the lead time $\varphi_{gl} = L(g,l) - \lfloor L(g,l)\rfloor$. The FLT model replaces (4) in model (1)-(6) with the following constraints:

$$\varphi_g X_{g,t-\lceil L(g)\rceil} + I_{g,t-1} - I_{gt}^{(\varphi_g)} - B_{g,t-1} + B_{gt}^{(\varphi_g)} = \varphi_g D_{gt}, \qquad g \in G, t = 1,\dots,T \quad (7)$$

$$Y_{gtl} = \varphi_{gl}X_{g,t-\lceil L(g,l)\rceil} + \left(1-\varphi_{gl}\right)X_{g,t-\lfloor L(g,l)\rfloor}, \qquad g \in G, t = 1,\dots,T, l \in O(g). \quad (8)$$

Constraint set (7) supplements (3) by representing the FGI and backlog at a point in time $t - 1 + \varphi_g$ within period t with a rate change in output at the end of the line based on differing release quantities $X_{g,t-\lceil L(g)\rceil}$ and $X_{g,t-\lfloor L(g)\rfloor}$. The added decision variables $I_{gt}^{(\varphi_g)}$ and $B_{gt}^{(\varphi_g)}$ contribute to FGI and backlogging costs in the objective function by a fractional amount of φ_g while I_{gt} and B_{gt} are weighted by $(1-\varphi_g)$. Constraints (8) represent the input-output relationship with fractional LTs. Initial WIP $W_{g,0,l}$ consumes capacity and is processed at operation $m \in O(g)$ in period t, if $t - 1 < L(g,m) - L(g,l-1) \le t$.

The following additional notation is required for the ACF formulation:

Sets and indices

$C(k)$: index set of the line segment used to approximate the CF for work center k

$K(g,l)$: work center where operation l of product g can be performed

Decision variables

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

Z_{gtl}^{k}: fraction of output from work center k allocated to operation l of product g in period t

Parameters

μ_k^n: intercept of segment n of the CF for work center k

β_k^n: slope of segment n of the CF for work center k.

The WIP balance (2), the input-output (4), and the capacity constraints (5) in model (1)-(6) are replaced by the following constraints that explicitly represent the CT behavior of the work centers:

$$W_{g,t-1,l} + X_{gtl} - Y_{gtl} = W_{gtl}, \quad g \in G, t = 1,\dots,T, l \in O(g) \tag{9}$$

$$\alpha_{gl} Y_{gtl} \le \mu_k^n Z_{gtl}^k + \beta_k^n (X_{gtl} + W_{g,t-1,l}), \quad g \in G, t = 1,\dots,T, l \in O(g), k \in K(g,l), n \in C(k) \tag{10}$$

$$\sum_{g \in G, l \in O(g,k)} Z_{gtl}^k = 1, \quad k \in K, t = 1,\dots,T \tag{11}$$

$$X_{gtl}, W_{gtl}, Z_{gtl}^k \ge 0, \quad g \in G, t = 1,\dots,T, l \in O(g), k \in K(g,l).$$

Constraints (9) represent the WIP balance for each operation. The CF constraint set (10) limits the output based on the available workload in units of time. The Z_{gtl} 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. The output allocation among operations is modeled by constraints (11). The CFs are fitted to empirical data from simulation as described by Kacar et al. (2013).

3.2 LT Estimation

Planning formulations with fixed exogenous LTs are dependent on appropriate LT estimates for their parameterization to achieve high-quality release schedules. While simulation models can be used to obtain accurate estimates, we focus on LT estimation based on historical data for the purpose of achieving a better understanding of the relationship between exogenous LTs and fab performance. LTs directly influence the expected FGI and backlog values in a planning model. For the FLT formulation, we derive FGI and backlog values at $t = L(g)$ in the planning window using constraints (7), (3), and (8):

$$\begin{aligned} I_{g,\lceil L(g)\rceil}^{(\varphi_g)} - B_{g,\lceil L(g)\rceil}^{(\varphi_g)} &= I_{g,\lfloor L(g)\rfloor} - B_{g,\lfloor L(g)\rfloor} + \varphi_g X_{g,0} - \varphi D_{g,\lceil L(g)\rceil} \\ &= I_{g0} - B_{g0} + \textstyle\sum_{\tau=1}^{\lfloor L(g)\rfloor} Y_{g\tau} + \varphi_g X_{g,0} - \sum_{\tau=1}^{\lfloor L(g)\rfloor} D_{g\tau} - \varphi D_{g,\lceil L(g)\rceil} \\ &= I_{g0} - B_{g0} + \textstyle\sum_{\tau=1}^{\lfloor L(g)\rfloor} X_{g\tau-\lfloor L(g)\rfloor} + \varphi_g X_{g,1-\lceil L(g)\rceil} - \sum_{\tau=1}^{\lfloor L(g)\rfloor} D_{g\tau} - \varphi D_{g,\lceil L(g)\rceil} \\ &= I_{g0} - B_{g0} + \textstyle\sum_{l \in O(g)} W_{g,0,l} - \sum_{\tau=1}^{\lfloor L(g)\rfloor} D_{g\tau} - \varphi D_{g,\lceil L(g)\rceil}. \end{aligned} \tag{12}$$

In (12), we use the assumption that initial WIP for product g and operation l represents lots released during the time interval $(-L(g,l), -L(g,l-1)]$. As newly released lots are delayed as specified in (8), they will have to compensate for this LT-dependent FGI and backlog in addition to covering subsequent demands. Furthermore, the LT estimate determines the set of operations at which both the initial WIP and newly released lots will consume capacity in each period. Overall, more capacity is consumed with lower LTs. This limits the WIP of product g to approximately $L(g)$ times its maximum possible TH at the bottleneck work center.

We distinguish four different types of LT estimates in the context of rolling horizon planning, namely constant, current, moving average, and forecast. Constant LTs $LT^{\mathrm{con}}(g)$ are based on historical data and represent the expected long-term behavior of the production system given a stationary demand generation process and planned release quantities. They are derived by averaging the LTs of initial simulation runs under the desired conditions. Based on (12), we expect a build-up of FGI in periods of low demand and accumulation of backlog in periods of high demand. The estimation of current LTs is usually based on the CTs of recently completed lots. However, these are representative for a time span equal to the CT itself. To represent the system's behavior concerning product g completed in period $t \leq 0$ where $t = 0$ represents the current planning time, we sum the average CTs $\overline{CT}_{glt}$ for all operations $l \in O(g)$:

$$LT^{\mathrm{cur}}(g,t) \coloneqq \sum_{l \in O(g)} \overline{CT}_{glt}\,, g \in G, t \leq 0.$$

Current LTs reflect the most recently observed state of the system, but they do not necessarily give better estimates for future CTs which depend on future release decisions. They may also fluctuate more strongly between periods. The consequences of this behavior on the system's performance are yet unknown. More stable LT estimates can be estimated by moving averages of the current LTs over the past H periods:

$$LT^{\mathrm{avg}}(g,t) \coloneqq \frac{1}{H}\sum_{\tau=t}^{t+1-H} LT^{\mathrm{cur}}(g,\tau)\,, g \in G, t \leq 0. \tag{13}$$

This allows for adjustments to long-term changes in demand and utilization without large short-term fluctuations. To get a forecast of future CTs, we propose a heuristic that incorporates information on initial FGI and backlog as well as the demand forecast. We assume that the planning formulation will ideally produce a release schedule satisfying all demand until the end of the planning horizon T with zero remaining FGI and backlog. This require an average TH of

$$TH_{gt}^{\mathrm{fct}} \coloneqq \frac{1}{T}(B_{gt} - I_{gt} + \sum_{\tau=1}^{T} D_{gt\tau}), g \in G, t \leq 0,$$

where $D_{gt\tau}$ denotes the demand forecast at the end of period t for product g and period $t+\tau$. The average throughput TH_{gt}^{avg} for the past H periods can be calculated as in (13). Based on these TH estimates, the expected work center utilization for both cases is given by:

$$U_{kt} \coloneqq {}^{1}/_{C_k} \sum_{g \in G, l \in O(g,k)} \alpha_{gl} TH_{gt}\,,\ k \in K, t \leq 0.$$

The constant work center capacities C_k can be replaced by time-dependent values. If detailed historical data for each operation is not available, the average CTs for each operation can be approximated using global flow factors:

$$FF_{gt} \coloneqq \frac{L(g,t)}{\sum_{l \in O(g)} \alpha_{gl}}\,, g \in G, t \leq 0.$$

Using flow factor values FF_{gt}^{avg} based on moving average LTs, we derive the LT forecast of each product by:

$$LT^{\mathrm{fct}}(g,t) \coloneqq \sum_{l \in O(g)} \left(\alpha_{gl} + \frac{U_{K(g,l),t}^{\mathrm{fct}}}{U_{K(g,l),t}^{\mathrm{avg}}} \left(FF_{gt}^{\mathrm{avg}} - 1\right)\alpha_{gl} \right), g \in G, t \leq 0.$$

Depending on the predicted work center utilization U_{kt}^{fct}, we get a point on the line through the processing time at a utilization of zero percent and the approximated moving average CT $FF_{gt}^{\mathrm{avg}}\alpha_{gl}$ at an

utilization equal to U_{kt}^{avg} for each operation and associated work center. The linear relationship is shown in Figure 1. As CTs grow nonlinearly with utilization, $LT^{\text{fct}}(g,t)$ can be considered a conservative estimate of changes in LT resulting in smooth transitions between periods. We expect this forecast to yield superior planning results compared to the other types of LT estimates in case of a demand impulse.

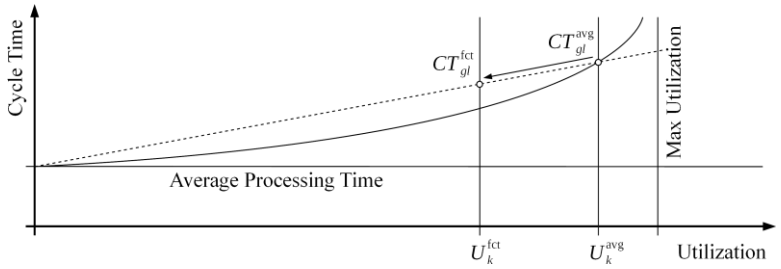

Figure 1: CT-utilization curve for a single operation and work center with CT forecast.

4 SIMULATION ENVIRONMENT

4.1 Simulation Infrastructure and Simulation Model

Our simulation experiments use the simulation infrastructure proposed by Ziarnetzky et al. (2015). Its center is a blackboard-type data layer in the memory of the simulation computer that resides between the planning and control level and the execution level, which is represented by a simulation model. The data layer contains important business objects such as routes, lots, and machines that are updated in an event-driven manner using notification functions of the simulation engine AutoSched AP 11.3. The planning level contains the different production planning models coded in ILOG CPLEX and the C++ programming language. The infrastructure allows for a rolling horizon approach where the simulation stops at the beginning of each planning epoch, i.e. when a new period starts, and instantiates a new planning instance based on feedback from the simulation model. This planning instance is then solved, and the release schedules are transferred to the control level where the specified number of lots of each product are released uniformly over the new period.

The MIMAC I simulation model (Fowler and Robinson 1995) is used in the simulation experiments. The model contains semiconductor manufacturing characteristics such as batch processing machines that can process several lots at the same time on a single machine, sequence-dependent setup times, exponentially distributed machine breakdowns, and operators. The model contains more than 200 machines organized into 69 work centers, with the planned bottleneck at the stepper work center, two products, each with more than 200 process steps, and uses First Out First In (FIFO) dispatching.

4.2 Demand Generation

Mean demand values that lead to the desired target bottleneck utilization (BNU) levels are determined using simulation. Demand in each period for each product is normally distributed, and the same mean demand is set for the two products to obtain mean BNU levels of 90%. For demand scenarios with a surge impulse (SI), the mean demand for exactly one product is increased for a specific number of periods. The SI is described by an impulse size (IS) and an impulse length (IL). During the SI, the mean demand for the surge product is multiplied by the IS given as a decimal fraction. All demand realizations are generated based on the predetermined mean demand M_{gs} for each product g and simulation period s with variability determined by the coefficient of variation (CV). Demand is generated for each period of the entire simulation horizon according to:

$$D_{gs} \coloneqq M_{gs}\left(1 + r_{gs}\right), s = 1, \ldots, t_{s,\max},$$

where $t_{s,\max}$ denotes the length of the simulation horizon and r_{gs} is a realization of the normally distributed random variable $R_1 \sim N(0, \sigma^2)$ with $\sigma = CV = 0.25$. Since demand is forecast-based, a demand volatility of $\eta = 0.05$ is used to generate the demand values for each period and product along the planning window of the planning occurrence s as follows:

$$D_{gst} := \begin{cases} D_{gs}, & \text{if } t = 1 \\ D_{g,s+t-1}(1 + \eta \tilde{r}_{gst}\sqrt{t-1}), & \text{if } t = 2, \ldots, T, \end{cases}$$

where $\tilde{r}_{gst}$ is a realization of the random variable $R_2 \sim N(0,1)$ and T the length of the planning window. Ten independent instances are generated for each demand scenario.

5 SIMULATION EXPERIMENTS

5.1 Design of Experiments

The experiments are designed to achieve a better understanding of the behavior of a wafer fab under high utilization when confronted with an SI. Different planning approaches are used to generate weekly release schedules based on the current state of the wafer fab and the demand forecast. Besides the simple backwards termination approach (REF), we are using two planning formulations with fixed LTs (SRD, FLT) and one with workload-dependent LTs (ACF). The fixed LT models are parameterized with four different types of LT estimates, three of which consider updated historical and forecast data available at the time of planning. For LT^{avg} the moving average time window is set to $H = 12$ periods. The SI scenarios have an impulse size of 133% and length of 12 periods (is133il12) or an impulse size of 200% and length of 4 periods (is200il4). The expected additional demand is the same in both scenarios, but distributed over a different time interval. The SI is applied to the demand for the second product starting in period s =21. For each demand scenario, ten independent demand realizations are considered. Each demand realization is simulated for twenty independent replications. The design of experiments is summarized in Table 1.

Table 1: Design of experiments for surge and overload analysis.

Factor	Level	Count
Planning approach	REF, SRD, FLT, ACF	4
Lead time estimate (SRD, FLT)	constant, current, moving average, forecast	4
Demand scenario	is133il12, is200il4	2
Demand realizations		10
Simulation replications		20
Total simulation runs		4000

The simulation replications are initialized from previously generated WIP snapshots and run for a total of 104 weeklong periods. A new release schedule is computed at the beginning of each period based on demand forecasts for 12 weeks. Three additional demand periods are added based on the average values of the previous three periods to account for end of horizon effects. Each combination of demand realization and simulation replication uses a unique WIP snapshot taken after one year of simulation time with a constant release rate corresponding to the specified BNU. The unit cost for WIP, FGI, and backlog are set to $\omega_{gt} = 60$, $h_{gt} = 10$, and $b_{gt} = 90$, whereas a unit revenue of 450 is assumed for the profit calculation. Data on cost, profit, and fab performance are gathered for individual periods during simulation. Performance measures for each period are averaged over all demand realizations and independent simulation replications since we are interested in analyzing the transient behavior of the wafer fab.

5.2 Results

The realized costs and profit for demand scenario is133il12 are given in Table 2. Although the SRD model considers finite capacities, it performs worse than the simple backward planning approach REF when parameterized with constant LTs. The LTs are not updated during the SI and rounded down by the SRD model. As indicated by (12), underestimated LTs result in wrong assumptions regarding FGI and backlog levels within the planning window, causing release schedules to be insufficient to satisfy given demands. Compared to the constant LTs, the periodically updated moving average and forecast estimates reduce the backlog by 16.3% and 29.3%, respectively. This comes at the cost of higher WIP levels. Current LTs produce similar FGI and backlog costs as the moving averages, but increase the WIP cost even further. The non-smooth change of LT values results in less stable release schedules and reduced system performance.

Table 2: Cost and profit values for SI with settings is133 and il12.

Planning approach	WIP cost	FGI cost	BLG cost	Total cost	Revenue	Profit
REF	1,406,127	13,128	427,683	1,846,938	3,042,293	1,195,354
SRD (constant)	1,304,272	982	568,893	1,874,147	3,027,384	1,153,237
SRD (current)	1,465,893	2,907	477,693	1,946,493	3,032,768	1,086,276
SRD (movavg)	1,351,194	2,943	476,225	1,830,362	3,032,462	1,202,100
SRD (forecast)	1,432,784	4,209	402,283	1,839,275	3,035,041	1,195,766
FLT (constant)	1,344,804	10,278	324,064	1,679,146	3,045,227	1,366,081
FLT (current)	1,447,064	13,514	310,255	1,770,833	3,048,721	1,277,888
FLT (movavg)	1,387,470	13,754	260,263	1,661,487	3,047,951	1,386,464
FLT (forecast)	1,409,065	13,826	228,259	1,651,149	3,049,511	1,398,362
ACF	1,466,486	19,848	196,089	1,682,423	3,058,191	1,375,768

Fully utilizing the fractional LT estimates allows for a much more accurate prediction of output quantities over time. Hence, the release schedules optimized by the FLT formulation substantially reduce the combined FGI and backlog costs for all LT estimates compared to SRD. FLT with forecast-based LT estimates produces the highest profit values. While profit values are slightly worse, ACF achieves the lowest combined FGI and backlog costs, being 10.8% below the forecast-based FLT variant. The results for an SI with settings is200il4 given in Table 3 are similar to those in Table 2.

Table 3: Cost and profit values for SI with settings is200 and il4.

Planning approach	WIP cost	FGI cost	BLG cost	Total cost	Revenue	Profit
REF	1,433,915	12,843	482,619	1,929,377	3,041,964	1,112,587
SRD (constant)	1,306,215	1,072	592,178	1,899,465	3,027,002	1,127,536
SRD (current)	1,471,129	2,967	503,371	1,977,467	3,032,071	1,054,604
SRD (movavg)	1,352,879	3,055	493,121	1,849,054	3,031,657	1,182,602
SRD (forecast)	1,448,134	4,729	425,590	1,878,452	3,035,862	1,157,410
FLT (constant)	1,346,834	10,277	349,194	1,706,305	3,046,165	1,339,860
FLT (current)	1,451,037	13,444	352,016	1,816,497	3,046,995	1,230,498
FLT (movavg)	1,388,343	14,044	284,551	1,686,938	3,049,153	1,362,215
FLT (forecast)	1,414,747	14,181	254,688	1,683,616	3,050,408	1,366,793
ACF	1,466,972	20,246	218,983	1,706,200	3,058,700	1,352,499

Total costs are slightly higher, with the total cost increasing more than twice as much for REF as for any other planning approach. Forecast-based LT estimates seem to be a slightly less robust to the shorter and more sizable SI compared to moving averages or the workload-based LTs of the ACF model.

5.3 Transient Analysis

We perform a transient analysis focusing on the evolution of WIP, FGI, and backlog over time in response to the SI. Figure 2 shows the development of these measures as the sum over both products for FLT parameterized with the different LT estimates given an SI with settings is133il12. The periods affected by the SI are indicated by the grey background between periods 21 and 32. Increased demand values appear in the demand forecast for the first time in period 10. In response, WIP levels increase sharply and remain elevated or continue to increase up until and during the SI.

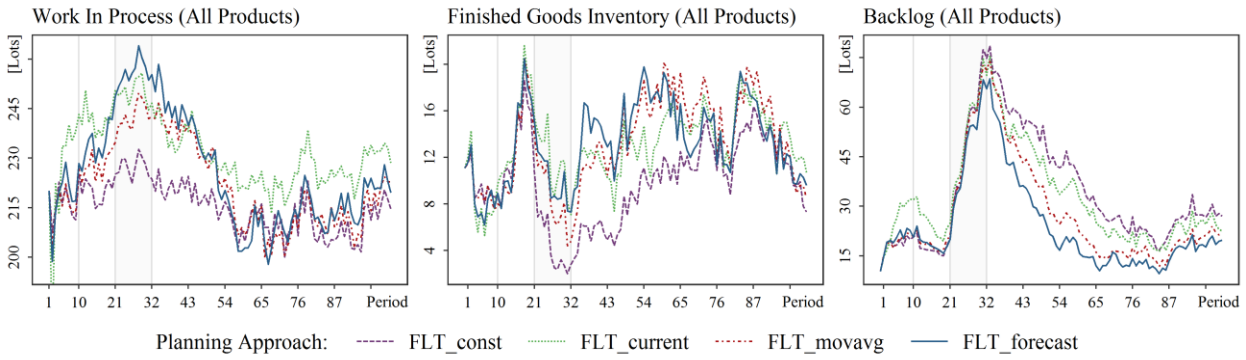

Figure 2: WIP, FGI, and backlog for the FLT model given a SI with settings is133il12.

Constant LTs limit the WIP, so FGI is depleted and backlog builds more quickly during the SI. To return to the previous average backlog level after the SI, the system needs almost another year. The LT forecast causes the largest increase in WIP after period 10. As a result, backlog increases less and is reduced more quickly among all LT estimates. Outside of the SI, current LTs result in the highest average WIP values. This appears to be due to a self-reinforcing effect where a high LT leads to high release quantities, which in turn increase the LT for the next planning period. Analyzing this SI scenario further, we compare the performance of all four planning approaches, only using forecast LT estimates for SRD and FLT. The WIP, FGI, and backlog values are shown in Figure 3.

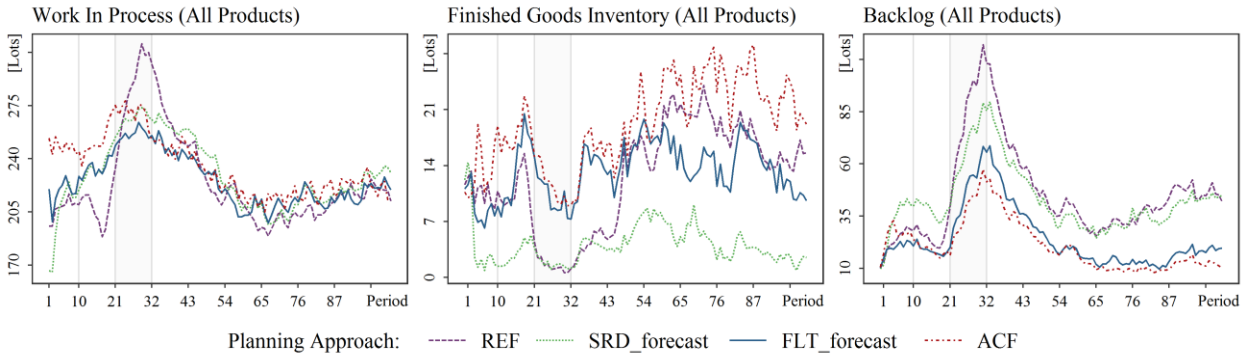

Figure 3: WIP, FGI, and backlog for the four different planning approaches using forecast LT estimates for the fixed LT models given a SI with settings is133il12.

The REF approach only considers demand forecasts for the periods explicitly specified by the constant integer LTs. Thus, WIP levels stay flat until $\lfloor LT^{\mathrm{con}}(g) \rfloor$ periods before the SI and increase sharply to the highest level among all approaches shortly before its end. Because of the limited FGI build-up and congestion caused by the high release quantities at the beginning of the SI, backlog reaches the highest

values among all approaches as well. The SRD formulation consistently lags demands due to the rounding down of LTs. Backlog during the SI rises less than with the REF approach, as the model accounts for limited capacities and ramps up WIP as a reaction to the demand forecasts beginning in period 10. ACF increases the WIP at the start of the simulation to a much higher value than all other planning approaches. Once TH has caught up with the initial increase in backlog due to congestion, it retains the highest levels of FGI and the lowest levels of backlog during the entire simulation horizon. This can be attributed to the ability of ACF to consider the cost advantage of building up FGI over production at higher utilization with nonlinearly increasing WIP in later periods. In contrast, formulations with exogenous LTs like SRD and FLT are unable to deliberately build-up FGI unless bottleneck utilization is expected to reach 100%.

6 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this paper, we compared production planning models that differ in how they anticipate the shop-floor behavior with product mix changes. Inspired by the previous study of Dümmler (2000), we tested a single SI. We demonstrated by designed simulation experiments that the ACF model with workload-dependent lead times outperformed the other formulations for combined FGI and backlog cost. The FLT formulation, that uses fractional exogenous LTs is able to generate higher profits if parameterized with appropriate LT estimates for each planning occurrence.

There are several directions for future research. First, it is desirable to repeat the study for simulation models from the SMT2020 testbed (Kopp et al. 2020) that are better representations of modern wafer fabs. As a second direction, it is interesting to use demand that is generated based on MMFE (Heath and Jackson 1994, Norouzi and Uzsoy 2014). Moreover, we are interested in studying the impact of production control strategies, i.e. dispatching rules, on the wafer fab performance during product mix changes. Preliminary simulation experiments indicate that the applied dispatching rule is important.

ACKNOWLEDGMENTS

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 this financial support.

REFERENCES

Asmundsson, J. M., R. L. Rardin, C. H. Turkseven, and R. Uzsoy. 2009. "Production Planning with Resources Subject to Congestion". *Naval Research Logistics* 56(2):142-157.

Asmundsson, J. M., R. L. Rardin, and R. Uzsoy. 2006. "Tractable Nonlinear Production Planning Models for Semiconductor Wafer Fabrication Facilities". *IEEE Transactions on Semiconductor Manufacturing* 19(1):95-111.

Dümmler, M. A. 2000. "Analysis of the Instationary Behavior of a Wafer Fab During Product Mix Changes". In *Proceedings of the 2000 Winter Simulation Conference*, edited by J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 1436-1442. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Dümmler, M., and O. Rose. 2000. "Analysis of the Short Term Impact of Changes in Product Mix". In *Proceedings of the International Conference on Modeling and Analysis of Semiconductor Manufacturing (MASM 2000)*, 133-138.

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

Gangsterer, M. 2015. "Aggregate Planning and Forecasting in Make-to-order Production Systems". *International Journal of Production Economics* 170:521-528.

Heath, D. C., and P. L. Jackson. 1994. "Modeling the Evolution of Demand Forecasts with Applications to Safety Stock Analysis in Production Distribution Systems". *IIE Transactions* 26(3):17-30.

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 Manufacturin*g 29(2): 153-167.

Kopp, D., M. Hassoun, A. Kalir, and L. Mönch. 2020. "SMT2020 - A Semiconductor Manufacturing Testbed". *IEEE Transactions on Semiconductor Manufacturing* 33(4):522-531.

Leachman, R. 2001. "Semiconductor Production Planning". In *Handbook of Applied Optimization*, edited by P. Pardalos, and M. Resende, 746-762, New York: Oxford University Press.

Metters, R. 1997."Production Planning with Stochastic Seasonal Demand and Capacitated Production". *IIE Transactions*, 29(11):1017-1029.

Metters, R. 1998. "General Rules for Production Planning with Seasonal Demand". *International Journal of Production Research* 36(5): 1387-1399.

Missbauer, H., and R. Uzsoy. 2011. "Optimization Models of Production Planning Problems". In *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*, edited by K. G. Kempf, P. Keskinocak, and R. Uzsoy, 437-508. New York: Springer.

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. 2018a. "A Survey of Semiconductor Supply Chain Models Part I: Semiconductor Supply Chains and Strategic Network Design". *International Journal of Production Research* 56(13):4524-4545.

Mönch, L., R. Uzsoy, and J. W. Fowler. 2018b. "A Survey of Semiconductor Supply Chain Models Part III: Master Planning, Production Planning, and Demand Fulfillment". *International Journal of Production Research* 56(13):4565-4584.

Nag, P. K., and W. Maly. 1995. "Cost of "Ad Hoc" Wafer Release Policies". In *Proceedings of IEEE/UCS/SEMI International Symposium on Semiconductor Manufacturing, September 17th to 19th, Austin TX, USA*, 97-102.

Norouzi, A., and R. Uzsoy 2014. "Modeling the Evolution of Dependency between Demands, with Application to Production Planning*". IIE Transactions* 46(1):55-66.

Rose, O. 1998. "WIP Evolution of a Semiconductor Factory After a Bottleneck Workcenter Breakdown". In *Proceedings of the 1998 Winter Simulation Conference*, edited by D. J. Medeiros, E. E Watson, J. S. Carson, and M. S. Manivannan, 997-1003. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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, 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., L. Mönch, and R. Uzsoy. 2018. "Rolling Horizon Production Planning for Wafer Fabs with Chance Constraints and Forecast Evolution". *International Journal of Production Research* 56(18): 6112-6134.

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.

REHA UZSOY is Clifton A. Anderson Distinguished Professor in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He holds BS degrees in Industrial Engineering and Mathematics and an M.S. in Industrial Engineering from Bogazici University, Istanbul, Turkey. He received his Ph.D. in Industrial and Systems Engineering in 1990 from the University of Florida. His teaching and research interests are in production planning, scheduling, and supply chain management. He was named a Fellow of the Institute of Industrial Engineers in 2005, Outstanding Young Industrial Engineer in Education in 1997, and has received awards for both undergraduate and graduate teaching. He can be reached by email at ruzsoy@ncsu.edu.