Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

SIMULATION-BASED PERFORMANCE ASSESSMENT OF PRODUCTION PLANNING FORMULATIONS FOR SEMICONDUCTOR WAFER FABRICATION

Timm Ziarnetzky

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

N. Baris Kacar

Operations Research Consulting Group

SAS Institute
Cary, NC 27513, USA

Lars Mönch

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

Reha Uzsoy

Edward P. Fitts Department of Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 276-7906, USA

ABSTRACT

In this paper we compare two production planning formulations in a rolling horizon setting. The first is based on fixed lead times that are a multiple of the period length, while the second uses non-linear clearing functions. A scaled-down simulation model of a wafer fab is used to assess the performance of the two formulations. We examine the impact of the planning window and period length on the performance of the production planning formulations. The performance advantage of clearing functions that is observed in a static setting can be also observed in a rolling horizon setting.

1 INTRODUCTION

Manufacturing integrated circuits on silicon wafers is one of the most complex manufacturing processes in use today. A diverse product mix that changes over time, reentrant process flows, many different machines with quite different performance characteristics and a high number of lots (the basic units of production control in semiconductor manufacturing) are typical for this type of production (Mönch et al. 2013). Production planning in these manufacturing facilities, known as wafer fabs, tries to fulfill demand by planning the quantities of material released into the fab over time to minimize costs.

The flow of material through capacity-constrained production facilities involves substantial delays due to queueing for congested resources. The average cycle time of the overall process can be of the order of several weeks in advanced wafer fabs (Mönch et al. 2013). Therefore, production planning formulations for wafer fabs must explicitly consider these delays. The estimates of cycle times used in production planning formulations are called lead times. Production planning formulations based on fixed lead times are discussed by Johnson and Montgomery (1974), Missbauer and Uzsoy (2010), and Voss and Woodruff (2003), among others.

Many planning systems in current use are based on the widely used Manufacturing Resource Planning (MRP II) approach (cf. Hopp and Spearman 2008), which uses deterministic lead time estimates as

exogenous parameters. The estimation of lead times for use in production planning models is far from trivial. Underestimating lead times will cause work to be released too late, resulting in underutilized resources and late delivery to customers. Overestimating lead times, on the other hand, will cause high work in progress (WIP) levels with the associated inventory costs, limiting the firm's ability to react to demand changes in a timely manner. Long lead times may also result in additional costs due to higher safety stock levels. We know from queueing theory and simulation models that cycle times increase in a nonlinear manner with increasing resource utilization. The resource utilization, however, is determined by the work release decisions that are a result of the planning process. Due to this circularity, cycle times need to be treated as an output of the planning process, rather than an input.

Researchers have recently begun to address the problem of production planning with workload-dependent lead times (cf. Asmundsson et al. 2006, Pahl et al. 2007, Asmundsson et al. 2009, Missbauer and Uzsoy 2010, Kacar et al. 2013a, among others). In this paper, we assess the performance of a production planning model with fixed lead times and a second model with workload-dependent lead times under a wide range of operating conditions by simulating the execution of the release schedules determined by the models in a rolling horizon setting. Little is known with respect to the behavior of production planning models with workload-dependent lead times in a rolling horizon setting.

The remainder of this paper is organized as follows. The problem setting is described in Section 2. This includes a discussion of related work. We then discuss the simulation environment used in course of the simulation experiments in Section 3. The results of the performed simulation experiments are reported in Section 4.

2 PROLEM SETTING

2.1 Planning Formulations

We consider a finite time horizon of length T divided into discrete periods of equal length. The objective of the models is to determine the amount of each product to release into the wafer fab in each period so as to minimize the costs caused by these releases. Multiple machine types with limited capacity organized in work centers are considered. The first linear programming (LP) formulation that assumes fixed lead times is given as follows:

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
$O(k)$:	set of all operations 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}:	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*

Parameters

h_{gt} : unit FGI holding cost for product *g* in period *t*

b_{gt} : unit backlogging cost for product *g* in period *t*

ω_{gt} : unit WIP 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 of operation *l* of product *g*

$L(g,l)$: estimated time elapsing from the release of the raw material of product *g* to the completion of the operation *l* of product *g*.

The first model can be stated as follows:

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

subject to

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

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

$$Y_{gtl} = X_{g,t-\lfloor L(g,l)\rfloor}, \quad \text{for all } g \in G, t = 1,\ldots,T, l \in O(g) \tag{4}$$

$$\sum_{g\in G}\sum_{l\in O(k)}\alpha_{gl}Y_{gtl} \le C_k, \quad \text{for all } k \in K, t = 1,\ldots,T \tag{5}$$

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

The objective (1) to be minimized is the sum of WIP, inventory, 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. Constraint set (3) represents FGI material balance at the end of the line. Constraints (4) define the relation between the time a lot of product *g* is released into the wafer fab and its completing processing at operation *l* of product *g.* As soon as a lot is processed at a given operation, it becomes available to the next operation on its routing. Constraint set (5) ensures that the total time required to process all operations at each work center in a given period *t* does not exceed the time available at that work center. The model assumes that an operation consumes capacity in the period that it is processed. Finally, constraints (6) ensure nonnegativity of the decision variables.

Model (1)-(6) incorporates lead time estimates. Let $L(g,l)$ be a fractional lead time estimate for operation l of product g. We compute $L(g,l)$ by the recursion:

$$L(g,l) := L(g,l-1) + FF_g\alpha_{gl}, \text{ for all } g \in G, l \in O(g), \tag{7}$$

where $L(g,0) := 0$. Here, FF_g denotes the flow factor of product *g*, defined as the ratio of the average time required for material started into the process to become available as FGI to the sum of the processing times of all its operations. FF_g values are obtained from long simulation runs for a given bottleneck utilization. Since we estimate the integer lead times by rounding down the fractional estimates obtained from simulation, we refer to this model as the Simple Rounding Down (SRD) model (Kacar et al. 2012). In contrast to conventional fixed lead time LPs in the literature, the SRD model contains WIP variables and WIP balance constraints to account for WIP present in the wafer fab at the beginning of the planning window. An appropriate initialization of the initial WIP in the planning formulation is important. For this reason, the initial WIP available at a given operation at the start of period 1 is treated as initial inventory

that must be consumed before any newly released material can be processed. The lead times for this material are also modified to represent the time required for the material to transition from their initial location to the subsequent operations on their routings.

Clearing functions (CFs) relate the expected output of a production resource in a planning period to some measure of the expected workload over that period. Early CF models had some difficulty in dealing with multiple products because it may be possible to create capacity for one product by holding WIP of another. To address this problem, Asmundsson et al. (2009) propose the Allocated Clearing Function (ACF) formulation where the output of the production resource is estimated using an aggregate workload measure and then allocated to individual products. The following additional notation is required for the ACF formulation:

Sets and indices

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

$K(l)$: work centers where operation l 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 objective function of the ACF formulation and the FGI material balance constraints are the same as in the SRD model. The WIP balance constraints (2), the fixed lead time constraints (4), and the capacity constraints (5) are replaced by new constraints that explicitly represent the cycle time behavior of the work centers. We obtain the following additional constraints:

$$W_{g,t-1,l} + X_{gti} - Y_{gti} = W_{gtl}, \quad \text{for all } g \in G, t = 1,\ldots,T, l \in O(g) \tag{8}$$

$$\alpha_{gl} Y_{gtl} \le \mu_k^n Z_{gtl}^k + \beta_k^n \alpha_{gl} \left(X_{gtl} + W_{g,t-1,l}\right), \quad \text{for all } g \in G, t = 1,\ldots,T, l \in O(g), k \in K(l), n \in C(k) \tag{9}$$

$$\sum_{g \in G, l \in O(k)} Z_{gtl}^k = 1, \quad \text{for all } k \in K, t = 1,\ldots,T \tag{10}$$

$$X_{gtl}, W_{gtl}, Z_{gt}^k \ge 0, \quad \text{for all } k \in K, g \in G, t = 1,\ldots,T, l \in O(g). \tag{11}$$

Constraints (8) ensure the WIP balance at each work center. In constraints (9), the CF relates the expected output of each work center in a period to the planned load of the work center in that period. The output allocation among operations is modeled by constraints (10). The 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. In the ACF formulation, any initial WIP at an operation is included in the argument of the CF that determines the amount of output the work center can produce in a given period. We refer the reader to Asmundsson et al. (2009) for more details of the ACF formulation.

We fit the CFs to empirical data obtained from a simulation model of the manufacturing system under consideration similar to the approach described in Kacar et al. (2013a). The simulation model is used to

collect observations of $Y_{kt} := \sum_{g \in G, l \in O(k)} \alpha_{gl} Y_{gtl}$, $X_{kt} := \sum_{g \in G, l \in O(k)} \alpha_{gt} X_{gtl}$, and $W_{k,t-1:} := \sum_{g \in G, l \in O(k)} \alpha_{gl} W_{g,t-1,l}$ for each period t and each work center k. The functional relationship between releases, initial WIP, and output Y_{kt} is established by dividing the resource load axis into two intervals containing an equal number of data points and fitting separate linear functions to the data in each segment using the linear regression function in SAS-OR. The third segment has a slope of zero and an intercept equal to the work center's theoretical capacity limit.

2.2 Related Work

Leachman (2001) proposes several high-fidelity production planning formulations for wafer fabs that consider fractional lead times. Model formulations based on integer-valued fixed lead times and CFs are proposed and compared in Asmundsson et al. (2006, 2009), Kacar et al. (2012), and Kacar et al. (2013a). The ACF formulations outperform the fixed lead time-based formulations for a wide range of situations in a static setting. Kacar et al. (2013b) find that fractional lead time-based formulations significantly outperform the SRD formulation in a static setting, yielding performance comparable to that of ACF. Häussler (2014) investigates the impact of the period length on the performance of ACF formulations.

Rolling horizon planning in supply chains is an important recent research topic (Sahin et al. 2013). Spitter (2005) discusses supply chain planning approaches in a rolling horizon setting. Master planning approaches for a simplified semiconductor supply chain are studied in a rolling horizon setting by Ponsignon and Mönch (2014). However, this paper assumes fixed lead times. A simulation-based framework is proposed that allows for executing the release schedules. An extensive literature review by Lin (2014) has studied the issue of schedule stability or nervousness, the repeated changes in planned quantities due to the replanning that occurs in the rolling horizon environment.

Rolling horizon approaches using CFs are rarely discussed in the literature. Stampfer et al. (2013) consider a small hybrid flow shop containing nine work centers, while Lin (2014) considers a single stage production system. Process conditions like reentrant flows or batching as found in wafer fabs are not taken into account in both papers. Orcun and Uzsoy (2011) examine the performance of rolling horizon methods using fixed lead times and CFs in a simple serial supply chain, and find that the different planning models lead to quite different dynamic behavior in the supply chain.

2.3 Problem Formulation

It is known from the literature (Ponsignon and Mönch 2014, among others) that rolling horizon approaches allow a more realistic performance assessment of planning approaches. In some situations, the advantage of optimization-based approaches is much smaller in a rolling horizon setting than in static settings. As pointed out by Sahin et al. (2013), we need a better understanding of rolling horizon methods in order to achieve planning stability without compromising the quality of the production plans.

In the present paper, we examine whether the advantage of the ACF formulation over SRD persists in a rolling horizon setting under different demand settings with forecast errors. We consider the expected profit as the main performance measure in this paper. Simulation experiments with a scaled-down simulation model of a wafer fab are performed to address this research question. We also investigate the impact of the planning window, the number of periods considered in the optimization model solved in each period, and the length of a planning period on the performance of the ACF formulation.

3 SIMULATION ENVIRONMENT

3.1 Simulation Framework

The simulation infrastructure we use for evaluating the ACF and SRD formulations consists of planning, control, and execution levels. It is based on the framework for simulation-based performance assessment

proposed by Ponsignon and Mönch (2014). A blackboard-type data layer in the memory of the simulation computer is the center point of the infrastructure. It is between the planning and execution level. The execution level is provided by a simulation model. A stop and go approach is taken, i.e., the simulation stops to compute a production plan that is determined using feedback from the simulation up to that point in time, and the plan is then transformed into a release schedule. The simulation then proceeds to implement this schedule until the next production plan has to be computed along the simulation timeline in a rolling horizon setting.

The production planning models and corresponding algorithms are implemented in the highest level of the infrastructure. The time between two consecutive planning occurrences is called the re-planning interval. For simplicity, we use a re-planning interval of one period. Demand fulfillment functionality is implemented to update the realized inventory and backlog values between two consecutive planning occurrences. Realized inventory, backlog, and WIP quantities from the execution level are stored in a blackboard-type data layer. The updates from the execution level at each planning occurrence are aggregated by the control level and incorporated as parameters into LP models. The LP models are generated based on the information stored in the data layer at the beginning of each planning occurrence on the planning level. The production planning algorithms determine production plans that are translated into lot release schedules by the control level, i.e., the lots to be released within a period are distributed uniformly over the period. Finally, the lot starts are triggered on the execution level and the processing of the corresponding lots is carried out. The infrastructure is coded in the C++ programming language, and the ILOG CPLEX libraries are used to solve the LP models, while AutoSched AP is used as the simulation engine. The overall architecture is depicted in Figure 1.

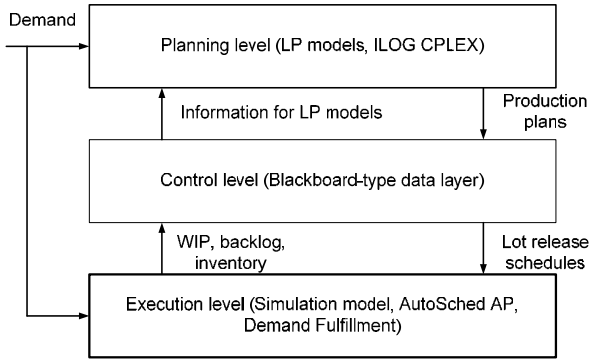

Figure 1: Architecture of the simulation infrastructure.

3.2 Simulation Model

The scaled-down representation of a wafer fab described by Kayton et al. (1997) is used to simulate the execution of the release schedules. Reentrant process flows, batch processing machines, unreliable machines, and multiple products with different process flows are included in the model based on typical attributes of a real-world wafer fab. The model consists of eleven work centers, each with one machine except for work center 4 that has two machines. First in First out (FIFO) dispatching is used at all work centers. Three products, each with a different number of process steps, are considered. The same lognormal distributed processing times are used for all products and process steps at each machine.

Instantaneous material transfer between successive process steps is assumed. Work centers 1 and 2 are batch processing machines with a minimum and maximum batch number of two and four lots, respectively. Any mix of lots can be processed as a batch. The expensive and scarce resources for the photolithography process in semiconductor manufacturing usually constitute the bottleneck in the fab. The bottleneck machines on the process flows of product 1 and 2 can be found at work center 4, while

that of product 3 can be found at work center 11. The batch processing machines in front of the bottleneck machines add variability to the arrival pattern of lots to be processed. Variability in the system is caused by Gamma distributions for the failures of the unreliable machines derived from work centers 3 and 7, which can cause starvation at the bottleneck machines.

4 SIMULATION EXPERIMENTS

4.1 Design of Experiments

According to Subsection 2.3, we expect that the performance of both the ACF and SRD formulations depends on the planning window and the period length. We investigate a period length of one day and of seven days, respectively. The cycle times of the products range from around 16 hours to more than one day. Two different planning windows that are a multiple of the period length are studied for each period length. Prior to generating demand instances, simulation runs are performed to obtain mean demand values for one period that lead to different bottleneck utilization levels. This setting is called stationary load. A product mix of 3:1:1 is taken for the mean demand values. In the level load (ll) scenarios, we generate normally distributed demand for each product in each period to obtain the desired mean bottleneck utilization levels of 70% and 90% over the simulation horizon. This results in negatively correlated demand where the product mix in each period varies, but the average resource utilization in each period remains constant at either 70% or 90%, depending on the specified utilization level.

We divide the simulation horizon into three-week subintervals in the time-varying (tv) demand load scenarios. For the scenarios with 90% average utilization, the utilization for each subinterval is selected to be either 85% or 95% with equal probability. The demand for each product is then set to achieve this level of bottleneck utilization. This leads to an average utilization level of 90% across all the periods. For the case of 70% average utilization, we set the utilization levels for the subintervals to be either 60% or 80% with equal probability to obtain an average utilization level of 70% across all periods. In this case the demands of the products are positively correlated. The demand scenarios are generated based on the level of resource utilization and the degree of variability that is determined by the coefficient of variation (CV). Demand for product g is generated for each period of the entire simulation horizon according to:

$$d_k^{(g)} := M_k^{(g)}(1 + r_k), k = 1, \ldots, t_{s,\max}, \tag{12}$$

where $t_{s,\max}$ denotes the length of the simulation horizon, $M_k^{(g)}$ the mean demand for product g in period k and r_k is a realization of the normally distributed random variable $R_1 \sim N(0, \sigma^2)$ with $\sigma = CV$. Because the demand is based on forecast, 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 n as follows:

$$D_{nt}^{(g)} := \begin{cases} d_n^{(g)}, & \text{if } t = 1 \\ d_{n+t-1}^{(g)}\left(1 + \eta\, \tilde{r}_{nt}\sqrt{t}\right), & \text{if } t = 2, \ldots, t_{\max}, \end{cases} \tag{13}$$

where $\tilde{r}_{nt}$ is a realization of the random variable $R_2 \sim N(0,1)$ and $t_{\max}$ the length of the planning window. To examine the effects of system variability, long and short machine failure durations are considered. The Gamma distributed Mean Time to Failure (MTTF) and Mean Time to Repair (MTTR) values in the model of Kayton et al. (1997) are used to represent the short failure case. The long failure case is obtained by doubling the MTTR and the MTTF values yielding the same average availability as in the short failure scenario. Different CFs are fit for each level of failure duration. Five independent instances are generated for each demand scenario and ten replications are performed for each instance to obtain statistically significant results. The design of experiments is summarized in Table 1.

The simulation is run for $t_{s,\max} = 104$ and $t_{s,\max} = 182$ periods in case of a period length of one week and one day, respectively. The production plan is revised after a single period, i.e., only the first period is

implemented. Long simulations are executed at the beginning of the simulation experiments to take a snapshot of the location of lots in front of the machines. Based on this, an initial WIP distribution is chosen to reduce initialization effects.

Table 1: Design of experiments.

Factor	Level	Count
Planning model	ACF, SRD	2
Period length (days)	1, 7	2
Planning window (periods)	7, 15	2
Mean utilization over horizon	low, high	2
Utilization over time	stationary load, level load, time-varying load	3
CV	0.1, 0.25	2
Machine failures	short, long	2
Mean demand scenarios		5
Simulation replications per demand scenario		10
Total simulation runs		9600

The realized profit is used for performance assessment. The unit revenue value is 60, while the unit backlog, WIP, and inventory costs per week are 50, 35, and 15, respectively. All experiments are carried out on a computer with 3.6 GHz Intel Core(TM) i7-4790 CPU and 16GB RAM. The computing times for SRD are, depending on the planning window, up to 10 seconds, while the corresponding times for ACF are up to 100 seconds.

4.2 Simulation Results

We start by presenting the profit values obtained from the rolling horizon setting for SRD and ACF. The realized profit per period for different planning windows is depicted in Figure 2 for high mean utilization. Due to space limitations, we do not show the corresponding results for low mean utilization since the results are similar to the high utilization case. The left- and right-hand chart sets show the results for period lengths of one and seven days, respectively. For instance, the notation tv90-10-S describes the average profit value of all demand scenarios and replications for a utilization level of 90% with time-varying demand, a demand CV of 0.1, and short failure durations. The notation ACF 7 indicates that ACF is used as planning formulation and that the planning window is seven periods. Figure 2 shows that ACF outperforms SRD in many situations. The impact of the length of the planning window on the profit is limited, probably since re-planning occurs after each planning period. At the same time, the release quantities cannot be very different for a different length of the planning window because the release decisions are heavily influenced by the treatment of the initial WIP in the period.

The output-release relationship (4) and the capacity assignment of the SRD formulations are more accurate for short period lengths. The CF represents the expected output in a period, and this gets much more variable as the period gets shorter. Thus, the profit difference for ACF and SRD is smaller for a period length of one day. In addition, the shorter planning period leads to more frequent updating of information between planning and execution levels.

Cost is the major cause of the different profit values when changing the period length as seen from Figure 3, since all models are constrained to meet demand as far as possible. The bars are labeled consistently with Figure 2, with the addition of the name of the corresponding planning formulation. The results for period lengths of one and seven days are distinguished where the results for one day periods are to the left of the heavy vertical bar. The average backlog, inventory, and WIP costs per period of all instances over the demand scenarios, simulation replications, and planning windows are plotted.

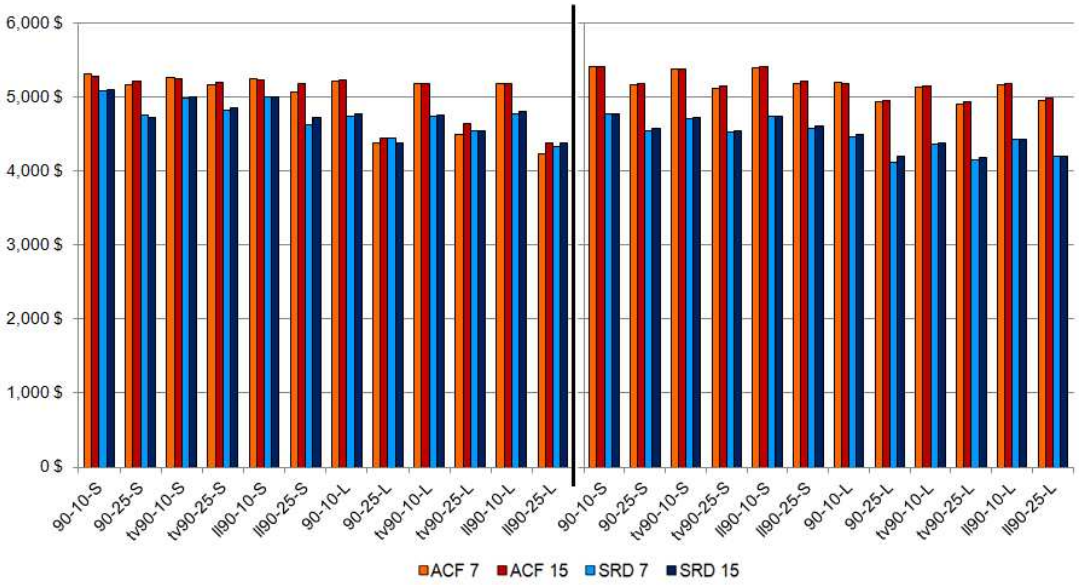

Figure 2: Profit comparison for different period lengths.

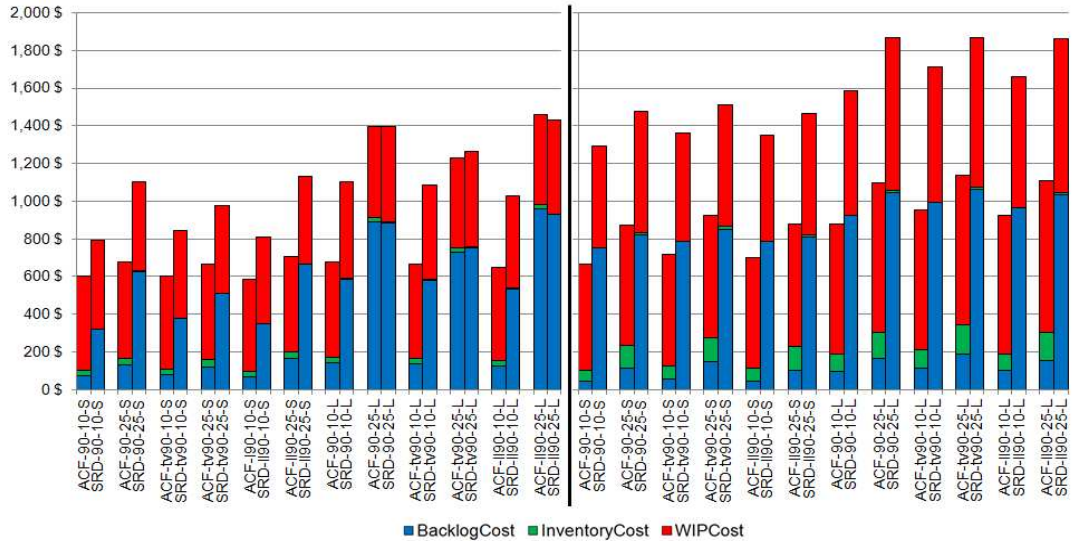

Figure 3: Cost distribution for different period lengths.

In the static setting, the SRD formulation suffers from regularly and significantly changing cycle times under time-varying demand pattern. The changes under level load-type demand are smoother and cause a better performance of the SRD formulation. Frequent updates from the execution level in the rolling horizon setting reduce the impact of the different demand patterns on the performance, unlike in the static setting where the demand pattern had a marked impact on the relative performance of the two models (Kacar et al. 2013a). A consistent observation from Figure 3 is that the ACF model holds both FGI and WIP, whereas the SRD model holds no FGI and incurs high backlogs. This suggests that SRD is overestimating the ability of the system to produce output in a given period of time, causing it to release work too late. While the ACF model holds somewhat higher WIP than SRD, the difference is not

extreme, and the cost of the additional WIP is more than offset by the reduction in backorder costs due to more timely release of work into the line. In the SRD model lead times are treated as deterministic exogenous parameters, but at high utilization levels small fluctuations in workload cause large changes in cycle time. The nonlinear CFs capture the workload-dependent lead times represented by output quantities and yield substantial improvements, in particular, in case of high utilization levels. Higher demand variability and longer failure durations result in lower profit and affect the difference between the results of the two planning formulations. The ratio of the average realized profit from the ACF model to those from the SRD model for a planning window of 15 days are summarized in Table 2. We see that improvements up to 19% are possible.

Table 2: Ratio of the realized profit for ACF relative to that from SRD.

Period length	**70-10-S**	**70-25-S**	**70-10-L**	**70-25-L**	**90-10-S**	**90-25-S**	**90-10-L**	**90-25-L**
	Stationary demand							
1 day	1.01	1.01	1.01	1.01	1.04	1.10	1.09	1.01
7 days	1.10	1.10	1.10	1.11	1.13	1.13	1.15	1.18
	Time-varying demand							
1 day	1.01	1.01	1.02	1.01	1.05	1.07	1.09	1.02
7 days	1.11	1.11	1.11	1.12	1.14	1.13	1.17	1.18
	Level load demand							
1 day	1.01	1.01	1.01	1.01	1.05	1.10	1.08	1.00
7 days	1.11	1.11	1.11	1.12	1.14	1.13	1.17	1.19

We compare the results of the different planning models for each demand realization using the Friedman test (Conover 1980) to assess the statistical significance of the results. While we do not present details of this analysis due to space limitations, in most experimental conditions the ACF model outperforms the SRD model at a significance level of 0.95.

5 CONCLUSIONS AND FUTURE RESEARCH

In this paper, we compared two different production planning formulations using a rolling horizon approach. The first formulation is based on fixed lead times that are a multiple of the planning period length, while the second formulation applies nonlinear CFs. Using a scaled-down simulation model of a wafer fab, we demonstrated that the advantage of the formulation with nonlinear CFs carries over to the rolling horizon setting. We also demonstrated that the period length has an impact on the performance of the production planning formulations with CFs, i.e., larger periods lead to an improved performance of the formulations based on CFs.

There are several directions for future research. First of all, we have to repeat the experiments from this paper for larger simulation models as in (Kacar et al. 2013a). Secondly, we are interested in using more advanced demand models, such as the Martingale Model of Forecast Evolution (MMFE) (cf. Heath and Jackson 1994, Chen and Lee 2009, Norouzi and Uzsoy 2014) in order to build planning models that can ensure a specified service level under demand uncertainty. Albey et al. (2014) present some initial results in this direction under a static planning environment. Thirdly, it is interesting to study the stability of production plans when demand uncertainty is taken into account in a rolling horizon setting. We expect that the stability of production plans can be increased by considering frozen periods or by using appropriate release change costs as proposed by Lin (2014).

ACKNOWLEDGMENTS

The research of Reha Uzsoy was supported by the National Science Foundation under Grant No. CMMI-1029706.

REFERENCES

Albey, E., A. Norouzi, K.G. Kempf, and R. Uzsoy. 2014. "Demand Modeling with Forecast Evolution: An Application to Production Planning." Technical Report, Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC.

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.

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* 19:95-111.

Chen, L., and H. L. Lee 2009. "Information Sharing and Order Variability Control Under a Generalized Demand Model." *Management Science* 55(5):781-797.

Conover, W. J. 1980. *Practical Nonparametric Statistics.* New York: John Wiley.

Häussler, S. 2014. "Comparison of Two Optimization based Order Release Models with Fixed and Variable Lead Times and an Empirical Validation of Metamodels of Work Centres in Order Release Planning." Ph.D. thesis, Department of Information Systems, Production and Logistics Management, University of Innsbruck.

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.

Hopp, W. J. and M. L. Spearman 2008. *Factory Physics: Foundations of Manufacturing Management*. Boston: Irwin/McGraw-Hill.

Johnson, L. A. and D. C. Montgomery 1974. *Operations Research in Production Planning, Scheduling and Inventory Control.* New York: John Wiley.

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 2013a. "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 2013b. "A Comparison of Production Planning Formulations with Exogenous Cycle Time Estimates Using a Large-Scale Wafer Fab Model." In *Proceedings of the 2013 Winter Simulation Conference*, edited by M. Kuhl, R. Pasupathy, S.-H. Kim, and A. Tolk, 3731-3744. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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):51-77.

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

Lin, P.-C. 2014. "Managing Release Changes in Rolling Horizon Production Planning." Ph.D. thesis, Edward P. Fitts Department of Industrial & Systems Engineering, North Carolina State University.

Missbauer, H., and R. Uzsoy 2010. "Optimization Models for Production Planning." 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.

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.

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

Orcun, S., and R. Uzsoy. 2011. “The Effects of Production Planning on the Dynamic Behavior of a Simple Supply Chain: An Experimental Study. “ In *Planning in the Extended Enterprise: A State of the Art Handbook*, edited by K.G. Kempf, P. Keskinocak, and R. Uzsoy, 43-80. Berlin: Springer.

Pahl, J., S. Voss, and D. L. Woodruff 2007. “Production Planning with Load Dependent Lead Times: An Update of Research.“ *Annals of Operations Research* 153:297-345.

Ponsignon, T., and L. Mönch 2014. “Simulation-based Performance Assessment of Master Planning Approaches in Semiconductor Manufacturing.“ *OMEGA* 46:21-35.

Sahin, F., A. Narayanan, and E. P. Robinson 2013. “Rolling Horizon Planning in Supply Chains: Review, Implications and Directions for Future Research.“ *International Journal of Production Research* 51(18):5413-5436.

Stampfer, C., S. Haeussler, and H. Missbauer. 2013. “The Impact of Foreknowledge of Demand in Case of Optimization-based Order Release Mechanism in Workload Control: A Simulation Study based on a Make-to-order Manufacturer.“ Technical Report, Department of Information Systems, Production and Logistics Management, University of Innsbruck.

Spitter, J. M. 2005. “Rolling Schedule Approaches for Supply Chain Operations Planning.“ Ph.D. thesis, Technische Universiteit Eindhoven.

Voss, S., and D. L. Woodruff (2003). *Introduction to Computational Optimization Models for Production Planning in a Supply Chain.* Berlin, New York: Springer.

AUTHOR BIOGRAPHIES

TIMM ZIARNETZKY is a Ph.D. student at the Chair of Enterprise-wide Software Systems, University of Hagen. He received M.S. degree in Mathematics from the Technical University Dortmund, Germany. His research interests include production planning and simulation-based production control. He can be reached by email at <Timm.Ziarnetzky@fernuni-hagen.de >.

NECIP BARIS KACAR is an Operations Research Specialist at the SAS Institute. He holds a Ph.D. degree in Industrial Engineering with Minor in Operations Research from the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University, and also holds a M.S. from the same university. He received a BS degree in Mechanical Engineering from Bogazici University, Istanbul, Turkey. His research interests are in production planning, supply chain management, inventory optimization and simulation based optimization. He can be reached via email at <Baris.Kacar@sas.com>.

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>.