Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

INTEGRATED PLANNING OF PRODUCTION AND ENGINEERING ACTIVITIES IN SEMICONDUCTOR SUPPLY CHAINS: A SIMULATION STUDY

Timm Ziarnetzky
Lars Mönch

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

Thomas Ponsignon
Hans Ehm

Infineon Supply Chain Innovation
Infineon Technologies AG
Am Campeon 1-15
Neubiberg, 85579, GERMANY

ABSTRACT

Running engineering lots is crucial to stay competitive in the semiconductor market. But production and engineering lots compete for the same expensive equipment. Therefore, considering them in an integrated way is desirable. In this paper, we propose two production planning formulations based on linear programming (LP) for a simplified semiconductor supply chain. The first planning model is based on reduced capacity for production due to engineering lots, while the second model directly incorporates engineering activities. Additional capacity is considered in the latter model due to learning effects that represent process improvements. Both planning models are based on exogenous lead times that are an integer multiple of the planning period length. We show by means of a simulation study for a simplified semiconductor supply chain that the integrated formulation outperforms the conventional one in a rolling horizon setting with respect to profit.

1 INTRODUCTION

Frequent engineering activities are required in semiconductor wafer fabrication facilities (wafer fabs) due to short product life cycles and the fierce competition in the semiconductor market (Mönch et al. 2018a). The different types of lots in wafer fabs, namely production lots and engineering lots, lead often to a situation where the production and engineering organization compete with each other for the scarce capacity of the expensive machines, operators, and engineering staff. Up to 30 percent of the overall capacity of a wafer fab can be consumed by engineering lots (Atherton and Atherton 1995; Leachman et al. 2002; Crist and Uzsoy 2011). It is pointed out by Chung and Huang (2002) that a correct modeling of the behavior of engineering lots is crucial to obtain correct cycle time estimates for wafer fabs.

There are only a few attempts to model competing production and engineering lots on the production control level. Tailored dispatching strategies for production and engineering activities are proposed by Crist and Uzsoy (2011) and to some extent also by Chung et al. (2015). However, decisions related to the allocation of resources to production and engineering activities are also possible on the production planning level. Thus, engineering activities need to be incorporated not only in production control procedures but at the same time also in production planning approaches. Despite of the obvious practical importance only little research is available for production planning approaches that take into account engineering activities (Kim and Uzsoy 2008; Kim and Uzsoy 2013). A detailed investigation of production planning models that incorporate engineering activities for a single wafer fab is conducted by Ziarnetzky and Mönch (2016). Learning effects are modeled that describe process improvement by performing engineering activities. Additional available capacity which results in increased future output is a result of learning. A rolling horizon setting is used to assess the performance of the different formulations. In the present paper, we are interested in extending these planning formulations to an entire

semiconductor supply chain. However, in addition to learning effects we also take into account that after performing a certain amount of engineering activities there is a reduced need for engineering staff which results in lower WIP cost. A preliminary version of the resulting models is already published in the extended abstract by Ziarnetzky et al. (2017). However, due to space limitations the planning models are not described in detail and the performance assessment results are only presented in a very aggregated way. In this paper, however, we present detailed versions of the models and complete simulation results.

This paper is organized as follows. We will describe the problem in the next section. This includes also a discussion of related work. We present the two LP formulations in Section 3. The supply chain simulation environment is discussed in Section 4. Moreover, the simulation results are analyzed and discussed in this section too. Conclusions and future research directions are provided in Section 5.

2 PROBLEM SETTING

2.1 Problem Statement

Different types of engineering activities exist in front-end (FE) and back-end (BE) facilities. The following types summarized in Table 1 are differentiated in this paper.

Table 1: Types of engineering lots.

Type	Description
Product Full (PF)	All process steps are performed that belong to the route of the engineering lot, but certain process steps of the lots have longer and more variable processing and setup times. Product development is the purpose.
Technology Partial (TP)	These engineering lots undergo only selected process steps for technology development in the metallization process that consists of alu-sputtering, litho-graphy, etching, and testing. The processing times are longer and more uncertain.
Equipment Verification (EV)	The functioning of a given machine is tested by running engineering lots. Only the process steps until visiting the steppers a second time are performed.
Equipment New (EN)	These engineering lots are used for testing the functioning of a new machine that has to be integrated onto the shop floor. Only the process steps until the second visit of the stepper work center are performed.

Note that we assume for the sake of simplicity that all engineering lots are unsalable. However, saleable samples exist in real-world semiconductor supply chains. In the present paper, we are interested in investigating a production planning formulation that deals with regular products and different types of engineering products as described in Table 1 in an integrated manner. We compare the integrated formulation that incorporates learning leading to an increase in available capacity with a conventional production planning formulation that is based on a static reduction of the available capacity for production. In contrast to previous research by Ziarnetzky and Mönch (2016) for a single wafer fab, we are interested in considering an entire semiconductor supply chain with several FE and BE facilities. The computational comparison of the two formulations has to be performed using a rolling horizon approach since we have to consider the execution level in a detailed manner. Moreover, in addition to the increased available capacity, we want to model the situation that the WIP cost is reduced after a certain number of engineering activities is performed. This setting again requires that a rolling horizon approach is taken.

2.2 Discussion of Related Work

We will discuss related work with respect to modeling engineering activities in production planning and control of semiconductor supply chains. The treatment of engineering activities in production control of wafer fabs is considered only in a few papers. Crist and Uzsoy (2012) propose several dispatching

strategies that take into account engineering lots. A scaled-down simulation model of a wafer fab is used to assess these strategies. A reservation-based dispatching strategy for engineering lots is proposed by Chung et al. (2015). The strategy takes into account the large variation in the processing times of the lots. Production planning formulations for both production and engineering activities based on clearing functions (CFs) are proposed by Kim and Uzsoy (2008), (2013) for the single- and multi-product case, respectively. The capacity allocation for production and engineering lots is explicitly modeled in these papers. A given fraction of the available capacity is reserved for engineering activities. Running engineering lots leads to an improved processing efficiency of production lots after a certain time delay due to learning. However, only a single-stage production system is assumed. Integrated production planning formulations based on exogenous lead times are proposed by Ziarnetzky and Mönch (2016) using the capacity allocation approach by Kim and Uzsoy (2008), (2013). The formulations are different with respect how engineering activities are incorporated and which information for the demand of engineering lots is available. In contrast to the present paper, only a single FE is considered. Rolling horizon experiments with the planning formulations show that the formulation where demand for engineering lots is only available for the first period outperforms the remaining integrated formulations and a conventional production planning model with static capacity reservation for engineering lots. Simulation optimization-based production planning formulations are studied by Manda and Uzsoy (2018) where new product transitions are taken into account. This stream of research is similar to the problem studied in the present paper since an exponential learning model is assumed and processing new products is similar to run engineering lots. However, only a single-stage production system is considered in this paper. The contribution of the present paper is twofold:

1. We extend the integrated formulation by Ziarnetzky and Mönch (2016) for a single FE to an entire semiconductor supply chain setting including both FE and BE facilities and perform simulation experiments in a rolling horizon setting for a simplified semiconductor supply chain.
2. The effect of changing WIP cost due to performing engineering activities on profit and cost is studied. This experiments are only possible if production planning and production control decisions are made together since dispatching policies are responsible for production and engineering lots after their release according to the planning decisions.

3 PRODUCTION PLANNING AND CONTROL

3.1 LP Formulations for Production Planning

We start by a conventional planning formulation where a fixed portion of the capacity is reserved for engineering activities. We assume that we have a planning window of T equidistant periods that are labeled by $t = 1,\ldots,T$. The following notation is used:

Sets and indices:

t: period index
g: production product index
j: facility index
k: work center index
l: operation index
G: set of all production products
F: set of all FE facilities
B: set of all BE facilities
$K^S(j)$: set of all work centers of stage S facility j, $S \in \{F,B\}$
$O^s(g,j)$: set of all operations of product g at stage S facility j, $S \in \{F,B\}$
$O^s(g,j,k)$: set of all operations of product g on machines of work center k of stage S facility j, $S \in \{F,B\}$

Decision variables:

$Y_{jt}^{S(g)}$: output of product g of stage S facility j from the last operation of its routing in period t, $S \in \{F,B\}$

$Y_{jtl}^{S(g)}$: quantity of product g of stage S facility j completing operation l in period t, $S \in \{F,B\}$

$X_{jt}^{S(g)}$; quantity of product g released into the first work center of stage S facility j in its routing in period t, $S \in \{F,B\}$

$W_{jt}^{S(g)}$: work in progress (WIP) of product g of stage S facility j at the end of period t, $S \in \{F,B\}$

$I_{jt}^{S(g)}$: finished goods inventory (FGI) of product g of stage S facility j at the end of period t, $S \in \{F,B\}$

$B_{jt}^{B(g)}$: distribution center (DC) backlog of product g of BE facility j at the end of period t

Parameters:

$h_{jt}^{S(g)}$: unit FGI holding cost for product g of stage S facility j in period t, $S \in \{F,B\}$

$\omega_{jt}^{S(g)}$: unit WIP cost for product g of stage S facility j in period t, $S \in \{F,B\}$

$b_{jt}^{B(g)}$: unit backlog cost for product g in period t of BE facility j

$u_{rst}^{S(g)}$: unit imbalance cost for product g between stage S facilities r and s in period t, $S \in \{F,B\}$

Λ_g: lot size relation between FE and BE lots of product g

D_{gt}: demand for product g during period t

C_{jkt}^{S}: available capacity of work center k of stage S facility j in period t, $S \in \{F,B\}$

$\alpha_{jl}^{S(g)}$: processing time of operation l of product g at stage S facility j, $S \in \{F,B\}$

$L_{jl}^{S(g)}$: lead time for product g from release of the raw material to the completion of operation l at stage S facility j, $S \in \{F,B\}$

q_{jk}^{F}: fraction of the capacity of work center k of FE facility j available for production activities.

The model can be formulated as follows:

$$\min \sum_{t=1}^{T} \sum_{g \in G} \left\{ \sum_{j \in F} \left[\omega_{jt}^{F(g)} W_{jt}^{F(g)} + h_{jt}^{F(g)} I_{jt}^{F(g)} \right] + \sum_{j \in B} \left[\omega_{jt}^{B(g)} W_{jt}^{B(g)} + h_{jt}^{B(g)} I_{jt}^{B(g)} + b_{jt}^{B(g)} B_{jt}^{B(g)} \right] \right. \tag{1}$$

$$\left. + \sum_{r,s \in F} u_{rst}^{F(g)} / 2 \left| X_{rt}^{F(g)} - X_{st}^{F(g)} \right| + \sum_{r,s \in B} u_{rst}^{B(g)} / 2 \left| X_{rt}^{B(g)} - X_{st}^{B(g)} \right| \right\}$$

subject to

$$W_{j,t-1}^{S(g)} + X_{jt}^{S(g)} - Y_{jt}^{S(g)} = W_{jt}^{S(g)},\ S \in \{F,B\},\ j \in S, t = 1,\ldots,T, g \in G \tag{2}$$

$$\sum_{j \in F} \left[\Lambda_g Y_{jt}^{F(g)} + I_{j,t-1}^{F(g)} - I_{jt}^{F(g)} \right] = \sum_{j \in B} X_{jt}^{B(g)},\ t = 1,\ldots,T, g \in G \tag{3}$$

$$Y_{jtl}^{S(g)} = X_{j,t-\left\lfloor L_{jl}^{S(g)} \right\rfloor}^{S(g)},\ S \in \{F,B\},\ j \in S, t = 1,\ldots,T, g \in G, l \in O^{S}(g,j) \tag{4}$$

$$\sum_{g \in G} \sum_{l \in O^{F}(g,j,k)} \alpha_{jl}^{F(g)} Y_{jtl}^{F(g)} \le q_{jk}^{F} C_{jkt}^{F},\ j \in F, t = 1,\ldots,T, k \in K^{F}(j) \tag{5}$$

$$\sum_{j \in B} \left[Y_{jt}^{B(g)} + I_{j,t-1}^{B(g)} - I_{jt}^{B(g)} + B_{jt}^{B(g)} - B_{j,t-1}^{B(g)} \right] = D_{t}^{(g)},\ t = 1,\ldots,T, g \in G \tag{6}$$

$$\sum_{g\in G}\sum_{l\in O^B(g,j,k)} \alpha_{jl}^{B(g)} Y_{jtl}^{B(g)} \le C_{jkt}^{B},\ j\in B, t=1,\ldots,T, k\in K^B(j) \tag{7}$$

$$X_{jt}^{S(g)}, Y_{jtl}^{S(g)}, Y_{jt}^{S(g)}, W_{jt}^{F(g)}, I_{jt}^{S(g)}, B_{jt}^{B(g)} \ge 0,\ S\in\{F,B\},\ j\in S, t=1,\ldots,T, g\in G, l\in O^S(g,j). \tag{8}$$

The objective function (1) to be minimized is the sum of FE FGI and WIP, BE backlog, FGI, WIP, and unit imbalance costs over all production products and periods in the planning window. The imbalance costs are used to penalize a situation where the number of started lots of the same product into two facilities of the same stage is quite different. A WIP-based formulation is used to represent the FE and BE WIP balance constraints (2), respectively and to include WIP cost in the formulation. The transfer from FE to BE including the lot size relation between FE and BE lots of each product is modeled by constraints (3), while the demand fulfillment is ensured by BE-related equations (6). Lead times are incorporated into the planning model by input-output relation constraints (4) for FE and BE, respectively, to describe the estimated final and die bank (DB) output quantities of the products for each period by the corresponding release quantities. The capacity consumption of each operation is assumed to take place at its completion. The capacity constraints (5) and (7) ensure that a maximum available capacity at each work center is respected. Equations (5) constrain the available FE capacity by a static capacity corridor reserved for engineering activities. Therefore, the finite capacity of the FE work centers is reduced to the fraction of the capacity that is allocated for production activities. Lots completing a certain operation become immediately available to the next operation on its routing. The decision variables are assumed to be non-negative by the constraint set (8).

The estimated cycle time L_{jgl}^{S} of product g at the stage S facility j elapsed from the release of the raw material to the completion of the operation l is determined by a recursive expression based on product-specific flow factors (see Kacar et al. 2013). The flow factor is defined as the ratio of the average time required for material started into the process to become available as FGI and the sum of the raw processing times of all its operations. Flow factors for each product are obtained from long simulation runs taking into account the desired bottleneck utilization (BNU). An appropriate initialization of the initial WIP in the planning formulation has to be considered. The release decisions for engineering lots are made based on an infinite capacity backward termination approach where the release of engineering lots into the wafer fab is based on the lead time and the demand of the corresponding product in each period. The estimated cycle time L_e for the last operation of the routing of an engineering product e is computed based on flow factors. A random number $r\in[0,1]$ is generated for each lot of the demand D_{et} for product e in period t to assign the lot release date into the first work center of its routing to a specific period. If r is smaller than the fractional part $L_e - \lfloor L_e \rfloor$ of the lead time L_e, the release period of the lot is $\max(t-\lfloor L_e\rfloor-1,1)$, otherwise it is $\max(t-\lfloor L_e\rfloor,1)$. The model (1)-(8) is called reduced capacity (RED) model. It takes into account engineering activities only indirectly by reducing the available capacity for each work center k to $q_k C_{jkt}^{B}$. The RED formulation is similar to the planning formulation proposed by Kacar et al. (2013). Next, we introduce an LP model that incorporates engineering activities as an integral part of the formulation. It is based on the insight that less demand information is available for engineering products since engineering lots are only requested on a short notice. Deterministic demand for engineering activities is only available for the current period. The demand uncertainty increases when future periods are taken into account. The integrated planning formulation differentiates between engineering products for TP and engineering activities for PF where the corresponding lots are processed in both the FE and the BE facilities. The following additional notation compared to the RED model (1)-(8) is required.

Sets and indices:

e:	engineering product index
E:	set of all engineering products
$\tilde{O}^{S(e)}(j)$:	set of all operations of product e at stage S facility j, $S\in\{F,B\}$

$\widetilde{O}^{S(e)}(j,k)$: set of all operations of product e on machines of work center k of stage S facility j, $S \in \{F,B\}$

Decision variables:

$\widetilde{Y}_{jtl}^{S(e)}$: quantity of product e completing its operation l of stage S facility j in period t, $S \in \{F,B\}$

$\widetilde{Y}_{jt}^{S(e)}$: output of product e in period t of stage S facility j from the last operation of its routing, $S \in \{F,B\}$

$\widetilde{X}_{jt}^{S(e)}$: quantity of product e released into the first work center of stage S facility j of its routing in period t, $S \in \{F,B\}$

$\widetilde{I}_{jt}^{F(e)}$: DB FGI of product e (product development) of FE facility j at the end of period t

$\widetilde{B}_{jt}^{F(e)}$: backlog of product e (technology development) of FE facility j at the end of period t

A_{jkt}^{F}: additional capacity of work center k of FE facility j available in period t induced by engineering activities

$\widetilde{B}_{jt}^{B(e)}$: DC backlog of product e (product development) at the end of period t of BE facility j

Parameters:

$\widetilde{h}_{jt}^{S(e)}$: unit FGI holding cost for product e of stage S facility j in period t, $S \in \{F,B\}$

$\widetilde{b}_{jt}^{F(e)}$: unit backlog cost for product e (technology development) of FE facility j in period t

$\widetilde{c}_{jt}^{S(e)}$: unit start cost for product e of stage S facility j in period t, $S \in \{F,B\}$

$\widetilde{b}_{jt}^{B(e)}$: unit backlog cost for product e (product development) in period t of BE facility j

$\widetilde{D}_{1}^{(e)}$: demand for product e during the first period

$\widetilde{M}_{jt}^{S(e)}$: minimum number of units of product e (PF or TP) to be completed in period t at stage S facility j, $S \in \{F,B\}$

$\widetilde{\alpha}_{jl}^{S(e)}$: processing time of operation l of product e at stage S facility j, $S \in \{F,B\}$

$\widetilde{L}_{jl}^{S(e)}$: lead time for product e from release of the corresponding material to the completion of operation l at stage S facility j, $S \in \{F,B\}$

$\widetilde{q}_{jk}^{F}$: fraction of capacity of work center k of FE facility j available for engineering activities

$d_{jkl}^{F(e)}$: time lag between engineering activities for operation l of product e at work center k of FE facility j and additional capacity of the same work center becoming available

γ_{jk}^{F}: time window where the additional capacity from engineering activities is available at work center k of FE facility j

U_{jk}^{F}: maximum additional capacity of work center k of FE facility j

$V_{jkl}^{F(e)}$: improvement rate by engineering activities for operation l of product e at work center k of FE facility j

V_{jk}^{F}: product of the maximum additional capacity of work center k of FE facility j and the improvement rate by engineering activities at the same work center, i.e. $V_{jk}^{F} \equiv U_{jk}^{F} V_{jkl}^{F(e)}$

$N^{F}(j,k)$: set of line segments used to approximate the additional capacity from the learning effect for work center k of FE facility j

μ_{njk}^{F}: intercept of line segment n for additional capacity at work center k of FE facility j due to engineering activities

β_{njk}^{F}: slope of line segment n for additional capacity at work center k of FE facility j due to engineering activities.

Due to space limitations we present only the model ingredients that are used in addition to model (1)-(8). The modified objective function (9) has to take into account the cost of performing engineering activities. It is given as follows:

$$\begin{aligned}
&\min \sum_{j\in F}\sum_{t=1}^{T}\left(\sum_{g\in G}\left[\omega_{jt}^{F(g)}W_{jt}^{F(g)}+h_{jt}^{F(g)}I_{jt}^{(g)E}\right]+\sum_{e\in E}\left[\tilde{c}_{jt}^{F(e)}\tilde{X}_{jt}^{F(e)}+\tilde{b}_{jt}^{F(e)}\tilde{B}_{jt}^{F(e)}\right]\right)+\\
&\sum_{j\in B}\sum_{t=1}^{T}\left(\sum_{g\in G}\left[\omega_{jt}^{B(g)}W_{jt}^{B(g)}+h_{jt}^{B(g)}I_{jt}^{B(g)}+b_{jt}^{B(g)}B_{jt}^{B(g)}\right]+\sum_{e\in E}\left[\tilde{c}_{jt}^{B(e)}\tilde{X}_{jt}^{B(e)}+\tilde{b}_{jt}^{B(e)}\tilde{B}_{jt}^{B(e)}\right]\right)+\\
&\sum_{r,s\in F}\sum_{t=1}^{T}\left[\sum_{p\in G\cup E}\frac{u_{rst}^{F(p)}}{2}\left|X_{pt}^{F(p)}-X_{qt}^{F(p)}\right|\right]+\sum_{r,s\in B}\sum_{t=1}^{T}\left[\sum_{p\in G\cup E}\frac{u_{rst}^{B(p)}}{2}\left|X_{pt}^{B(p)}-X_{qt}^{B(p)}\right|\right].
\end{aligned} \tag{9}$$

The following constraints (10)-(12) ensure that demand for engineering lots in the first period and the minimum number of engineering lots are reached in a given period. We obtain:

$$\sum_{j\in B}\left[\tilde{Y}_{j1}^{B(e)}+\tilde{B}_{j1}^{B(e)}-\tilde{B}_{j0}^{B(e)}\right]\geq\tilde{D}_{e1}, \qquad e\in E \tag{10}$$

$$\sum_{j\in B}\left[\tilde{Y}_{jt}^{B(e)}+\tilde{B}_{jt}^{B(e)}-\tilde{B}_{j,t-1}^{B(e)}\right]\geq\tilde{M}_{jt}^{B(e)}, \qquad t=2,\ldots,T, e\in E(PF) \tag{11}$$

$$\sum_{j\in F}\left[\tilde{Y}_{jt}^{F(e)}+\tilde{B}_{jt}^{F(e)}-\tilde{B}_{j,t-1}^{F(e)}\right]\geq\tilde{M}_{jt}^{F(e)}, \qquad t=2,\ldots,T, e\in E\setminus E(PF). \tag{12}$$

The next constraint set (13) models the transition from a FE to a BE facility for engineering activities of type PF, while constraint set (14) represents the input-output relationship if this is appropriate for the considered type of engineering activities:

$$\sum_{j\in F}\left[\Lambda_g\tilde{Y}_{jet}^{F}+\tilde{I}_{j,e,t-1}^{F}-\tilde{I}_{jet}^{F}\right]=\sum_{j\in B}\tilde{X}_{jet}^{B}, \qquad t=1,\ldots,T, e\in E(PF) \tag{13}$$

$$\tilde{Y}_{jtl}^{S(e)}=\tilde{X}_{j,t-\left\lfloor\tilde{L}_{jl}^{F(e)}\right\rfloor}^{S(e)}, \qquad j\in F, t=1,\ldots,T, e\in E, l\in\tilde{O}^{S(e)}(j), S\in\{F,B\}. \tag{14}$$

The constraint set (15a) models the learning effect as a concave, continuous, and non-decreasing function of the cumulative number of engineering lots processed on FE facility j work center k in a time window of length γ_{jk}^{F} where after a certain amount of elapsed time performing engineering activities leads to additional available capacity. A concave shape is chosen to mimic the improvement reduction that occurs if the number of performed engineering lots increases. Here, the expression A_{jkt}^{F} for the additional capacity due to the learning effect (15a) has to be piecewise linearized. This linearization is

modeled by (15b). The FE capacity constraints (5) from the RED model are replaced by constraints (16)-(17) with an upper bound on capacity available for production and engineering activities. The available FE capacity C^F_{jkt} is complemented by the constraints (16) that determine the additional available capacity caused by engineering activities. The amount of engineering activities is limited by constraint set (17). Constraint set (18) replaces the BE capacity constraint set (7). We have:

$$A^F_{jkt} = C^F_{jkt} \sum_{e\in E} \sum_{l\in \tilde{O}^{F(e)}(j,k)} U^F_{jk} \left[1 - \mathrm{e}^{-V^{F(e)}_{jkl} \sum_{\tau=t-d^{F(e)}_{jkl}}^{t-d^{F(e)}_{jkl}} \tilde{Y}^{F(e)}_{j\tau l}} \right] \qquad j\in F, t=1,\ldots,T, k\in K^F(j) \qquad (15a)$$

$$A^F_{jkt} \le \mu^F_{njk} + \beta^F_{njk} C^F_{jkt} V^F_{jk} \sum_{e\in E} \sum_{l\in \tilde{O}^{F(e)}(j,k)} \sum_{\tau=t-d^{F(e)}_{jkl}-\gamma^F_{jk}}^{t-d^{F(e)}_{jkl}} \tilde{Y}^{F(e)}_{j\tau l}, \qquad j\in F, t=1,\ldots,T, k\in K^F(j), n\in N^F(j,k) \qquad (15b)$$

$$\sum_{g\in G} \sum_{l\in O^{F(g)}(j,k)} \alpha^{F(g)}_{jl} Y^{F(g)}_{jtl} + \sum_{e\in E} \sum_{l\in \tilde{O}^{F(e)}(j,k)} \tilde{\alpha}^{F(e)}_{jl} \tilde{Y}^{F(e)}_{jtl} \le C^F_{jkt} + A^F_{jkt}, \qquad j\in F, t=1,\ldots,T, k\in K^F(j) \qquad (16)$$

$$\sum_{e\in E} \sum_{l\in \tilde{O}^{F(e)}(j,k)} \tilde{\alpha}^{F(e)}_{jl} \tilde{Y}^{F(e)}_{jtl} \le \tilde{q}^F_{jk} \left[C^F_{jk} + A^F_{jkt} \right], \qquad j\in F, t=1,\ldots,T, k\in K^F(j) \qquad (17)$$

$$\sum_{g\in G} \sum_{l\in O^{B(g)}(j,k)} \alpha^{B(g)}_{jl} Y^{B(g)}_{jtl} + \sum_{e\in E} \sum_{l\in \tilde{O}^{B(e)}(j,k)} \tilde{\alpha}^{B(e)}_{jl} \tilde{Y}^{B(e)}_{jtl} \le C^B_{jkt}, \qquad j\in B, t=1,\ldots,T, k\in K^B(j). \qquad (18)$$

Finally, the non-negativity of the additional decision variables is ensured by the constraint set:

$$\tilde{X}^{S(e)}_{jt}, \tilde{Y}^{S(e)}_{jtl}, \tilde{Y}^{S(e)}_{jt}, A^F_{jkt}, \tilde{B}^{S(e)}_{jt}, \tilde{I}^{F(e)}_{jt} \ge 0, S\in\{F,B\}, j\in S, t=1,\ldots,T, e\in E, l\in \tilde{O}^{S(e)}(j), k\in K^S(j). \qquad (19)$$

The resulting model (9)-(14), (15b)-(19) and (2)-(4), (7), (8) is an LP. It is abbreviated by simple rounding down (SRD) model. In the remainder of this paper we are interested in comparing the performance of the RED and SRD formulations in a rolling horizon setting where demand updates over time are taken into account.

3.2 Production Control Scheme

We apply dispatching strategies that differentiate between production and engineering lots. The four strategies due to Crist and Uzsoy (2011) are summarized in Table 2.

Table 2: Production control strategies.

Strategy	Description
Production-First (PF)	Production lots are always processed prior to any waiting engineering lot.
Engineering-First (EF)	Engineering lots are always processed prior to any waiting production lot.
Capacity-Allocation-to-Engineering (CAtE)	Capacity corridors of fixed length for engineering lots are used to process engineering lots. Only production lots are processed outside the corridor.
Change-over-Trigger (CoT)	A prescribed daily number of engineering lots is set as a threshold value. If this value is reached, this amount of engineering lots is processed on the machine. If only engineering lots wait for processing they will be processed without taking into account the threshold value.

The capacity corridors in the CAtE strategy are also called engineering intervals. Note that the CAtE strategy is a conventional approach while the more dynamic capacity allocation proposed by the integrated production planning formulation requires more sophisticated strategies such as EF, PF, and

CoT. Additional parameters are required to configure the CAtE and the CoT strategies. We have to set the length of the engineering interval ie and the periodically repeated start time of this interval for the CAtE. The remaining time is called the production interval. An integer-valued trigger threshold Δ must be set for the CoT strategy.

4 SIMULATION STUDY

4.1 Simulation Infrastructure and Supply Chain Simulation Model

The MIMAC 1 model (Fowler and Robinson 1995) provides a single FE facility in the simplified supply chain while the single BE facility Backend-I is taken from the supply chain testbed proposed by Ewen et al. (2017). Both models are modified to incorporate engineering process improvement activities. The different types of engineering lots are added based on the two products of the simulation models. A DB is located between FE and BE, and a DC is considered at the end of the supply chain. The submodels of the supply chain simulation model are publicly available at Testbed (2019). The simulation model is depicted in Figure 1. A rolling horizon approach is implemented that considers feedback from the simulation model when generating a planning instance at a new planning epoch. The realized backlog, FGI, and WIP and the fulfilled demand are updated between consecutive planning epochs in a blackboard-type data layer. Production plans are transformed into release schedules by a uniform distribution of the production lot release quantities over the period. Engineering lots are launched at the beginning of the periods. The simulation model is built in AutoSched AP. The dispatching strategies and production planning formulations are implemented in the C++ programming language using the customization functionality of the simulation tool. ILOG CPLEX is used to solve the LP models.

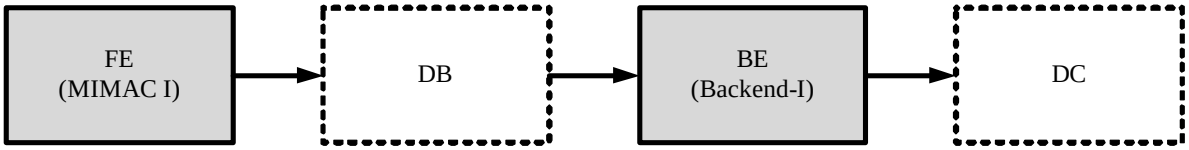

Figure 1: Simulation model.

The dispatching strategies from Subsection 3.2 are only applied to the planned bottleneck work center, i.e. to the steppers. The highest priority (HP) lot is processed first. The simulation model allows for capacity expansions due to performing engineering lots. The expansion is implemented by activating additional machines of the stepper work centers for a certain amount of time.

4.2 Design of Experiments

The goal of the simulation study is to compare the performance of the RED and the SRD formulation. Engineering activities have to be considered at the execution level to mimic an execution of planning decisions in the nodes of the supply chain. The profit, i.e. the difference of revenue and total costs, is used as performance measure. Total costs are the sum of backlog, inventory, FE and BE WIP costs for production lots and start and backlog costs for engineering. Revenue is only obtained by production lots. 10% of all WIP lots belong to engineering, namely 5% belong to TP, 3% to PF, and 1% each to EV and EN. Production lots have a regular priority, while engineering lots have a higher priority. Production lots have 48 wafers, while engineering lots have typically 24 wafer, but we have only two wafers per lot for EV. The FE unit revenue is 250. Moreover, we set $\tilde{c}_{1t}^{F(e)} = 20$, $\omega_{1t}^{F(g)} = 15$, $h_{1t}^{F(g)} = 15$, $b_{1t}^{B(g)} = 15$, and $b_{1t}^{F(e)} = 50$ for TP lots. Since we use $\Lambda_g = 3$, the remaining BE-related revenue and cost settings can be derived from the corresponding FE settings by taking only one third of them. We use $q_{1k}^{F} = 0.1$ and $\tilde{q}_{1k}^{F} = 0.2$. We set $d_{1kl}^{F(e)} = 1$ period, $\gamma_{1k}^{F} = 3$ periods, $U_{1k}^{F} = 0.1$, and $V_{1kl}^{F(e)} = 0.01$. The learning curve is approximated by three line segments. In addition, we set $\Delta = 5$ lots. Two engineering intervals per period

(week) are used. Each of them has a duration of $ie = 8.4$ hours. We also consider scenarios where the WIP cost is reduced as a result of running a certain amount of engineering lots. In the first setting, we have fixed WIP cost that is 40% of the original WIP cost, while the WIP cost decrease is low. In the second scenario the fixed WIP cost is 20% percent, while the decrease is steep (high). The former setting is abbreviated by low and the latter one by high. Desired BNU levels of 70% and 90% are applied. Normally distributed demand is considered. A coefficient of variation of $CV = 0.1$ represents the demand variability for the final demand given by

$$d_{g\tau} := M_{g\tau}\left(1 + r_\tau\right), \tau = 1,\ldots,t_{\max}, \tag{20}$$

where $M_{g\tau}$ is the mean demand for product g in period τ that leads to the desired BNU, $t_{\max}$ is the length of the simulation horizon, and r_τ is a realization of the normally distributed random variable $R_1 \sim N\left(0,\sigma^2\right)$ with $\sigma = CV$. Because the demand is based on forecast, we use a demand volatility of $\eta = 0.05$ to generate the demand for the planning instance of each planning epoch n as

$$D_{gt}^{(n)} := \begin{cases} d_{gn}, & \text{if } t = 1 \\ d_{g,n+t-1}\left(1 + \eta\, \tilde{r}_t^{(n)} t\right), & \text{if } t = 2,\ldots,T \end{cases}, \tag{21}$$

where T is the length of the planning window and $\tilde{r}_t^{(n)}$ a realization of the random variable $R_2 \sim N(0,1)$. Five different demand scenarios are considered for each BNU level. Ten independent simulation replications are performed for each scenario. The RED model is applied together with the CAtE strategy, whereas the combinations SRD + EF, SRD + PF, SRD + CoT are considered for the integrated formulation. We simulate $t_{\max} = 52$ periods with $T = 12$ and a period length of one week.

4.3 Simulation Results

We show the results of the simulation experiments in Table 3. Best results are marked in bold. Profit obtained by RED + CAtE is for both BNU settings the smallest. The SRD + EF and SRD + CoT slightly outperform the remaining SRD-dispatching strategy combinations in the case of a low and high BNU level, respectively.

Table 3: Realized production revenue and cost.

	BNU (%)	Revenue	FE WIP	Backlog	FGI	BE WIP	Engineering Cost	Total Cost	Profit
RED + CAtE	70	523,267	91,106	264,264	1255	34,045	21,546	412,216	111,051
SRD + EF	70	523,143	90,691	260,186	1202	33,780	10,384	396,242	**126,902**
SRD + PF	70	523,457	90,609	261,531	1180	33,829	10,407	397,556	125,900
SRD + CoT	70	523,409	90,671	261,270	1176	33,810	10,408	397,334	126,076
RED + CAtE	90	822,837	180,388	386,157	92	55,184	29,441	651,262	171,575
SRD + EF	90	825,568	171,320	375,133	53	54,870	16,462	617,838	207,731
SRD + PF	90	826,122	171,556	376,169	53	54,938	17,549	620,265	205,857
SRD + CoT	90	827,063	170,920	373,693	55	54,798	16,635	616,100	**210,963**

More engineering lots are released by the integrated approach. Hence, the engineering-related backlog costs are smaller and the profit increases. More production lots are finished when BNU=90% is considered. This is caused by the additional capacity due to running engineering lots. Overall, the

integrated formulation is beneficial. The simulation results for changing WIP costs are shown in Figure 2. Note that only simulation experiments for BNU=90% are conducted in this situation. We observe from Figure 2 that the advantage of the integrated formulation over the RED model carries over to the setting with changing WIP cost. However, the magnitude of improvement is smaller.

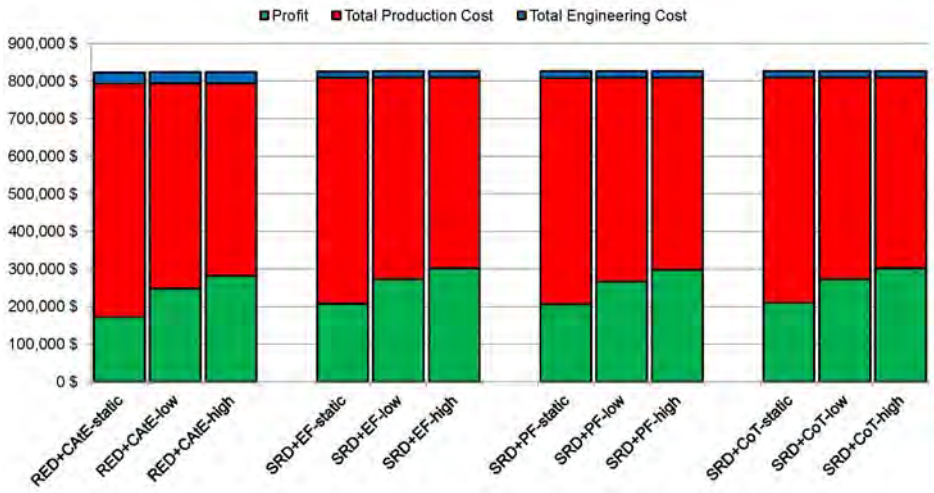

Figure 2: Simulation results for changing WIP costs.

5 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

An integrated planning formulation for both production and engineering activities was discussed in the present paper by extending the single FE planning formulation from Ziarnetzky and Mönch (2016) to the semiconductor supply chain level. The conventional formulation and the integrated one were assessed in a rolling horizon setting using a simulation model of a simplified semiconductor supply chain. The integrated formulation outperformed the conventional one with respect to profit under all experimental conditions.

There are several directions for future research. First of all, the planning models discussed in the present paper are based on the assumption of given, exogenous lead times that are an integer multiple of the period length, so-called fixed lead times. The assumption of fixed lead times is clearly not appropriate since the lead times are a result of the resource utilization which depends on the release decisions of a planning model (Mönch et al. 2018b). Therefore, integrated planning formulations with workload-dependent lead times, namely CFs, have to be proposed and tested. It is promising to apply the conic programming approaches used by Gopalswamy (2019) to avoid the linearization of constraint set (15a) in the SRD model. As a third direction of future research it might be interesting to try to extend the data-driven approaches proposed by Omar et al. (2017) to the present situation with engineering activities.

ACKNOWLEDGEMENTS

The research was partially supported by the iDev 4.0 project funded by the ECSEL Joint Undertaking (JU) under grant agreement No 783163. The authors gratefully acknowledge this financial support.

REFERENCES

Atherton, L. F. and R. W. Atherton. 1995. *Wafer Fabrication: Factory Performance and Analysis*. Norwell, MA: Kluwer Academic Publisher.

Chung, S. H. and H. W. Huang. 2002. "Cycle Time Estimation for Wafer Fab with Engineering Lots". *IIE Transactions* 34(2): 105-118.

Chung, Y. H., B. H. Kim, J. C. Seo, and S. C. Park. 2015. "Reservation Based Dispatching Rule for Wafer Fab with Engineering". 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, 2974-2982. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Crist, K. and R. Uzsoy. 2011. "Prioritizing Production and Engineering Lots in Wafer Fabrication Facilities: a Simulation Study". *International Journal of Production Research* 49(11):3105-3125.

Ewen, H., L. Mönch, H. Ehm, T. Ponsignon, J. Fowler, and L. Forstner. 2017. "A Testbed for Simulating Semiconductor Supply Chains". *IEEE Transactions on Semiconductor Manufacturing* 30(3):293-305.

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

Gopalswamy, K. 2019. *Production Planning with Clearing Functions: Data-driven Approaches and Conic Programming*. PhD Thesis, North Carolina State University, Raleigh.

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.

Kim, S., and R. Uzsoy. 2008. "Integrated Planning of Production and Engineering Process Improvement". *IEEE Transactions on Semiconductor Manufacturing* 21(3):390-398.

Kim, S., and R. Uzsoy. 2013. "Modeling and Analysis of Integrated Planning of Production and Engineering Process Improvement". *IEEE Transactions on Semiconductor Manufacturing* 26(3):414-422.

Leachman, R. C., J. Kang, and V. Lin. 2002. "SLIM: Short Cycle Times and Low Inventory in Manufacturing at Samsung Electronics". *Interfaces* 32 (1):61–77.

Manda, A. B., and R. Uzsoy, 2018. "Simulation Optimization for Planning Product Transitions in Semiconductor Manufacturing Facilities". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3470-3481. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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):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.

Testbed 2019. http://p2schedgen.fernuni-hagen.de/index.php?id=296, accessed 3rd May 2019.

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

Ziarnetzky, T., L. Mönch, T. Ponsignon, and H. Ehm. 2017. "Rolling Horizon Planning with Engineering Activities in Semiconductor Supply Chains". In *Proceedings CASE 2017*, August 20th -23rd, X'ian, China, 1024-1025.

AUTHOR BIOGRAPHIES

TIMM ZIARNETZKY received a diploma degree in mathematics from the Technical University Dortmund, Germany and a Ph.D. degree in Computer Science from the University of Hagen, Germany. His research interests include optimization and simulation of planning and control applications in complex production systems. His email address is timm.ziarnetzky@fernuni-hagen.de.

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

THOMAS PONSIGNON is a Senior Staff Supply Chain Engineer at Infineon Technologies in Neubiberg, Germany. He obtained master's degrees in Industrial Engineering from the EPF-Ecole d'Ingénieurs, Sceaux, France and the University of Applied Sciences, Munich, Germany and a Ph.D. in Computer Science from the University of Hagen, Germany. His email address is thomas.ponsignon@infineon.com.

HANS EHM is Lead Principal Supply Chain heading the supply chain innovation department at Infineon Technologies. He holds a diploma degree in Applied Physics from HS Munich and is M.S in Mechanical Engineering from Oregon State University. His email address is hans.ehm@infineon.com.