*Proceedings of the 2016 Winter Simulation Conference*
*T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.*

# SIMULATION-BASED OPTIMIZATION FOR INTEGRATED PRODUCTION PLANNING AND CAPACITY EXPANSION DECISIONS

Timm Ziarnetzky
Lars Mönch

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

## ABSTRACT

In this paper, we consider a simplified semiconductor supply chain that consists of a single front-end facility and back-end facility. We present a production planning formulation that is based on clearing functions. A cost-based objective function is considered. The minimum utilization of expensive bottleneck machines in the front-end facility is a parameter of the model. At the same time, the less expensive capacity of the back-end facility can be increased to reduce the cycle time in the backend facility. The release schedules obtained from the planning formulations are assessed using discrete-event simulation. An overall cycle time larger than a given maximum value is penalized. Simulated annealing is used to determine appropriate minimum utilization levels for the front-end bottleneck machines and appropriate capacity expansion levels for the back-end. The results of the computational experiments demonstrate that the profit can be increased while the maximum possible overall cycle time is not violated.

## 1 INTRODUCTION

Semiconductor supply chains are one of the most complex and challenging industrial environments that are in use today (cf. Chien *et al.* 2011). Semiconductor manufacturing starts with wafers, thin discs made of silicon. Up to 1000 chips can be manufactured on a single wafer by fabricating the ICs layer by layer in a wafer fabrication facility (wafer fab). Electrical tests that identify the individual dies that are likely to fail when packaged are performed in a probe/sort facility after the wafer fab step. Only the dies of appropriate quality will be put into a package. The wafer fab and the probe/sort stages are abbreviated by front-end (FE) steps. The probed wafers are then sent to an assembly facility where the good dies are put into an appropriate package. Finally, the assembled dies are sent to a test facility where they are tested in order to ensure that only high-quality chips are sent to customers (cf. Mönch *et al.* 2013). The assembly and the test stages are abbreviated by back-end (BE) steps.

A typical semiconductor supply chain (SC) consists of dozens of FE and BE facilities. The FE and BE facilities are decoupled by die banks (DBs) where wafers are stored before the BE processes start. The final chips are stored in distribution centers (DCs) before they are shipped to customers. Planning formulations for entire semiconductor supply chains (SCs) are challenging due to highly variable demand, reentrant flows in each single wafer fab, long cycle times of the lots, and a large number of different products. Only recently, researchers have began to propose corresponding SC-wide formulations (cf., for instance, Denton *et al.* 2006, Ponsignon and Mönch 2012, Lowe and Mason 2016).

In the present paper, we discuss an integrated production planning and capacity expansion problem for a simplified semiconductor SC that contains a single FE and BE facility. A maximum cycle time

constraint is taken into account. The decision variables that are chosen within a simulation-optimization scheme are the minimum utilization of the FE bottleneck and the amount of additional capacity of the BE facility to compensate the longer cycle times in the FE facility resulting from a higher bottleneck utilization.

The paper is organized as follows. The problem is discussed in Section 2. This section also provides a discussion of previous work. The integrated production planning and capacity expansion formulation is presented in Section 3. Moreover, the simulation-based optimization approach based on simulated annealing is sketched in this section too. Computational results are presented in Section 4. Conclusions and future research directions are discussed in Section 5.

## 2 PROBLEM SETTING AND DISCUSSION OF RELATED WORK

### 2.1 Problem

A model of a simplified SC that includes a single FE and a single BE facility is considered in the present paper. We assume a finite planning horizon of length $T$ that is divided into discrete equidistant periods. We are interested in determining a profit-maximizing release schedule of wafer quantities for the FE facility for several products and periods assuming a given deterministic demand. Moreover, since a small cycle time has a positive impact on yield and customer satisfaction (cf. Mönch *et al.* 2013), a maximum allowed cycle time is assumed. The FE-related part of the cycle time increases if more lots are released into the FE facility. At the same time, the BE-related part of the cycle time can be decreased by adding additional capacity to the BE facility. In contrast to the FE, where a capacity expansion often takes a long time and is extremely expensive, increasing capacity in BE facilities is less expensive and is also possible at short notice. Increasing the utilization of the expensive FE machines by releasing more lots might lead to a higher profit. Overall, we have to look for appropriate values for the minimum utilization of the FE bottleneck machines and the required capacity expansion of the BE facility.

### 2.2 Review of Related Work

We discuss related work with respect to capacity expansion decisions and production planning formulations for semiconductor SCs. Facility design formulations are proposed by Bard *et al.* (1999) assuming given deterministic demand. Queuing theory is used to estimate the waiting time of the lots in front of the work centers in a simulated annealing approach to minimize the average cycle time. Hopp *et al.* (2002) use an optimization approach based on queuing models to design a wafer fab in such a way that investment costs are minimized while a maximum allowed cycle time is maintained. A similar approach is taken by Sohn (2004) where the profit is maximized while a maximum cycle time constraint is ensured. However, in contrast to these papers, our problem is more operational. Therefore, a high-fidelity simulation model is more appropriate than the less detailed queuing models. Barahona *et al.* (2005) propose long-term capacity expansion models under demand uncertainty for a single wafer fab. But again, we consider an operational planning problem for an entire SC that requires different methods.

Various planning formulations for single wafer fabs and even entire semiconductor SCs are proposed in the literature. We refer, among others, to Denton *et al.* (2006), Ponsignon and Mönch (2012), and Lowe and Mason (2016) for entire SCs. The more detailed short-term production planning formulations differ in the way how the planned cycle time, i.e. the lead time, is modeled. Fixed integer lead time approaches where the lead time is an integer multiple of the period length (cf., for instance, Kacar *et al.* 2013) are differentiated from fraction lead time approaches (cf., for instance, Kacar *et al.* 2016). However, the main disadvantage of these approaches is, on the one hand, that the lead time is a parameter of the planning formulation. On the other hand, the cycle time depends in a nonlinear manner on the release schedule which is an output of the planning model. To avoid this circularity, more recently clearing function (CF)-based formulations are proposed (cf., for instance, Asmundsson *et al.* 2009, Kacar *et al.* 2013, Ziarnetzky *et al.* 2015). Here, concepts from queuing theory are used to formulate capacity

constraints in the models. No explicit lead time parameters are assumed in the resulting models. CF-based planning formulations are interesting for our problem because we cannot assume a fixed lead time for the FE facility since the minimum FE bottleneck machine utilization is a parameter of our planning formulation.

It seems promising to manage the search for appropriate parameters of production planning formulations by simulation-based optimization. This technique is, for instance, applied by Gansterer *et al.* (2014) to set appropriate lead time, safety stock, and lot size values in a production planning formulation for a make-to-order environment. A production planning problem with uncertain demand is solved for a single wafer fab based on simulation-based optimization by Liu *et al.* (2014). Overall, it seems reasonable for the problem addressed in the present paper to combine CF-based planning approaches with simulation-based optimization.

## 3 SIMULATION-BASED OPTIMIZATION APPROACH

### 3.1 Model Formulation

An allocated clearing function (ACF) and a fixed integer lead time (FLT) formulation are used for planning at the FE and BE facilities, respectively. Wafer fabrication and probe processes are integrated into the ACF formulation while assembly and final testing are embedded into the FLT formulation. A DB with instantaneous material transfer between the facilities from the FE to the BE facility is considered. The transportation time from the FE to the DB and from the BE to a DC is included in the ACF and FLT formulations, respectively. The integrated production planning formulation is given as follows:

Sets and indices

$G$: set of all products
$K^F$: set of all FE work centers
$K^B$: set of all BE work centers
$t$: period index
$g$: product index
$k$: work center index
$b$: FE bottleneck work center index
$l$: operation index
$O^F(g)$: set of all FE operations of product $g$
$O^B(g)$: set of all BE operations of product $g$
$O(k)$: set of all operations performed on machines of work center $k$
$C(k)$: set of indices denoting the line segment used to approximate the CF for FE work center $k$
$K(l)$: FE work centers where operation $l$ can be performed

Decision variables

$Y^F_{gtl}$: quantity of product $g$ completing its FE operation $l$ in period $t$
$Y^F_{gt}$: output of product $g$ in period $t$ from the last FE operation of its routing
$X^F_{gtl}$: quantity of product $g$ starting FE operation $l$ in period $t$
$W^F_{gtl}$: FE work in process (WIP) of product $g$ at operation $l$ at the end of period $t$
$Z^k_{gtl}$: fraction of output from FE work center $k$ allocated to operation $l$ of product $g$ in period $t$
$I^{DB}_{gt}$: DB finished goods inventory (FGI) of product $g$ at the end of period $t$

$Y_{gtl}^{B}$: quantity of product $g$ completing its BE operation $l$ in period $t$

$Y_{gt}^{B}$: output of product $g$ in period $t$ from the last BE operation of its routing

$X_{gt}^{B}$: quantity of product $g$ released into the first BE work center in its routing in period $t$

$W_{gt}^{B}$: BE WIP of product $g$ at the end of period $t$

$I_{gt}^{DC}$: DC FGI of product $g$ at the end of period $t$

$B_{gt}^{DC}$: DC backlog of product $g$ at the end of period $t$

Parameters

$\omega_{gt}^{F}$: unit FE WIP cost for product $g$ in period $t$

$h_{gt}^{DB}$: unit DB FGI holding cost for product $g$ in period $t$

$\omega_{gt}^{B}$: unit BE WIP cost for product $g$ in period $t$

$h_{gt}^{DC}$: unit DC FGI holding cost for product $g$ in period $t$

$b_{gt}^{DC}$: unit DC backlogging cost for product $g$ in period $t$

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

$C_k$: capacity of BE work center $k$ in units of time

$\tilde{C}_b$: capacity of FE bottleneck work center $b$ in units of time

$\alpha_{gl}$: processing time of operation $l$ of product $g$

$L(g,l)$: estimated time elapsing from the BE release of product $g$ to the completion of operation $l$ of product $g$

$\mu_k^n$: intercept of segment $n$ of the CF for FE work center $k$

$\beta_k^n$: slope of segment $n$ of the CF for FE work center $k$

$\Lambda_g$: lot size relation between FE and BE lots of product $g$, i.e. the FE lot size is $\Lambda_g$ times the BE lot size

$M$: minimum utilization of FE bottleneck work center $b$ (in percent).

The model can be formulated as follows:

$$\min \sum_{g\in G}\sum_{t=1}^{T}\left[\sum_{l\in O(g)}\omega_{gt}^{F}W_{gtl}^{F}+\omega_{gt}^{B}W_{gt}^{B}+h_{gt}^{DB}I_{gt}^{DB}+h_{gt}^{DC}I_{gt}^{DC}+b_{gt}^{DC}B_{gt}^{DC}\right] \quad (1)$$

subject to

$$W_{g,t-1,l}^{F}+X_{gtl}^{F}-Y_{gtl}^{F}=W_{gtl}^{F}, \quad \text{for all } g\in G,t=1,\ldots,T,l\in O^{F}(g) \quad (2)$$

$$\Lambda_g Y_{gt}^{F}+I_{g,t-1}^{DB}-I_{gt}^{DB}=X_{gt}^{B}, \quad \text{for all } g\in G,t=1,\ldots,T \quad (3)$$

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

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

$$\sum_{g\in G}\sum_{l\in O(b)}\alpha_{gl}Y_{gtl}^{F}\ge M\tilde{C}_b, \quad \text{for all } t=1,\ldots,T \quad (6)$$

$$W_{g,t-1}^{B}+X_{gt}^{B}-Y_{gt}^{B}=W_{gt}^{B}, \quad \text{for all } g\in G,t=1,\ldots,T \quad (7)$$

$$Y_{gt}^{B}+I_{g,t-1}^{DC}-I_{gt}^{DC}+B_{gt}^{DC}-B_{g,t-1}^{DC}\ge D_{gt}, \quad \text{for all } g\in G,t=1,\ldots,T \quad (8)$$

$$Y^B_{gtl} = X^B_{g,t-\lfloor L(g,l)\rfloor}, \qquad \text{for all } g \in G, t = 1,\ldots,T, l \in O^B(g) \tag{9}$$

$$\sum_{g\in G}\sum_{l\in O(k)} \alpha_{gl} Y^B_{gtl} \le C_k, \qquad \text{for all } k \in K^B, t = 1,\ldots,T \tag{10}$$

$$X^F_{gtl}, Y^F_{gtl}, W^F_{gtl}, Z^k_{gtl}, I^{DB}_{gt} \ge 0, \qquad \text{for all } k \in K^F, g \in G, t = 1,\ldots,T, l \in O^F(g) \tag{11}$$

$$X^B_{gt}, Y^B_{gtl}, Y^B_{gt}, W^B_{gt}, I^{DC}_{gt}, B^{DC}_{gt} \ge 0, \qquad \text{for all } g \in G, t = 1,\ldots,T, l \in O^B(g). \tag{12}$$

The objective function (1) to be minimized is the sum of inventory and WIP cost in the FE and the BE facility and the BE backlog cost over all products and periods. We abbreviate it by

$$C\left(W^F, W^B, B^{DC}, I^{DB}, I^{DC}\right) := \sum_{g\in G}\sum_{t=1}^{T}\left[\sum_{l\in O(g)} \omega^F_{gt} W^F_{gtl} + \omega^B_{gt} W^B_{gt} + h^{DB}_{gt} I^{DB}_{gt} + h^{DC}_{gt} I^{DC}_{gt} + b^{DC}_{gt} B^{DC}_{gt}\right]. \tag{13}$$

where the corresponding WIP, backlog, and inventory matrices are the arguments. We start by explaining the FE-related constraints. The WIP balance at each operation in the FE is ensured by constraints (2). The DB material balance constraints (3) describe how the BE release depends on the FE output. A single FE lot of product $g$ is split into $\Lambda_g$ BE lots. No backlog is allowed at the DB between FE and BE. Constraints (4) and (5) relate the expected FE output of each work center in a period to the planned load of the work center in that period and allocate it among the operations. The minimum FE bottleneck utilization in each period is ensured by constraints (6).

Next, the BE-related constraints are described. Constraints (7) represent the WIP balance of the BE for each product and each period. The FGI material balance constraints (8) allow release quantities that exceed the available demand. Constraints (9) express the relation between the time a lot of a given product is released into the BE and its completing processing at the specified operation of the product. The model takes into account the finite capacity of the machines. It is assumed that an operation consumes capacity in the period that it is processed. Therefore, the constraint set (10) ensures that the total time required to process all operations at each BE work center in a given period does not exceed the time available at that work center. The FE- and DB-related decision variables have to be non-negative according to constraint set (11), while the non-negativity of the BE- and DC-related decision variables is enforced by constraints (12). More details of related production planning formulations can be found in (Asmundsson *et al.* 2009, Kacar *et al.* 2013, Ziarnetzky *et al.* 2015).

The approach for fitting the CFs from empirical data obtained from simulation runs of the FE simulation model is the same as described in (Kacar *et al.* 2013). Moreover, the planning formulation (1)-(12) is equipped with lead time estimates for processing lots in the BE facility. Let $L(g,l)$ be a fractional lead time estimate for BE operation $l$ of product $g$. The quantity $L(g,l)$ is computed by the recursion:

$$L(g,l) := L(g,l-1) + FF_g \alpha_{gl}, \text{ for all } g \in G, l \in O^B(g). \tag{14}$$

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 BE process to become available as FGI to the sum of the processing times of all its BE operations. $FF_g$ values are obtained from long simulation runs using the simulation model of the simplified SC for a given bottleneck utilization and capacity expansion level. Larger flow factor values are obtained for higher bottleneck utilization levels and lower additional capacity.

### 3.2 Simulated Annealing Scheme

Simulation is useful to deal with the stochasticity of the base system of the SC. Therefore, we embed the planning formulation (1)-(12) into a simulation-based optimization approach. We are interested in determining appropriate values for the minimum utilization of the FE bottleneck, abbreviated by $M$, and

the additional BE capacity in percent, denoted by $a$. We start by introducing the following additional notation for formulating a new objective function:

$r_{gt}$: unit revenue for product $g$ in period $t$

$\beta$: penalty for exceeded cycle times

$\delta$: penalty for FE bottleneck utilization shortfall

$\lambda$: penalty for additional BE capacity

$C_{gt}$: realized average cycle time for product $g$ in period $t$

$C_{gt}^{*}$: upper limit of the average cycle time for product $g$ in period $t$

$U_{bt}$: realized utilization at FE bottleneck work center $b$ in period $t$ in units of time.

The objective function to be minimized is given as follows:

$$f(M,a) := \sum_{g\in G}\sum_{t=1}^{T} r_{gt}\tilde{Y}_{gt}^{B} - C\left(\tilde{W}^{F},\tilde{W}^{B},\tilde{B}^{DC},\tilde{I}^{DB},\tilde{I}^{DC}\right) - \frac{\beta}{T}\sum_{g\in G}\sum_{t=1}^{T}\left(C_{gt}-C_{gt}^{*}\right)^{+} - \frac{\delta}{T}\sum_{t=1}^{T}\left(M\tilde{C}_{b}-U_{bt}\right)^{+} - \lambda a\,. \quad (15)$$

where we set $x^{+} := \max(x,0)$. The tilde token is used to indicate that realizations from the base system, i.e. from the simulation, are taken as arguments. The objective function (15) is the difference of the realized profit and penalty terms for exceeding the maximum allowed cycle time and for falling short of the minimum FE bottleneck utilization to be reached in a period, and the capacity expansion costs in the BE facility. Note that the quantities $M$ and $a$ are parameters of the planning formulation, i.e., we use $C_k := (1+a)\hat{C}_k$ where $\hat{C}_k$ is the regular capacity of the BE work center $k$. Discrete-event simulation is used to evaluate the objective function (15). Note that the cycle time values and the FE bottleneck utilization values are taken from the base system, while the WIP, backlog, and inventory quantities are a result of the executed release schedules computed by the planning model (1)-(12). The overall setting is depicted in Figure 1.

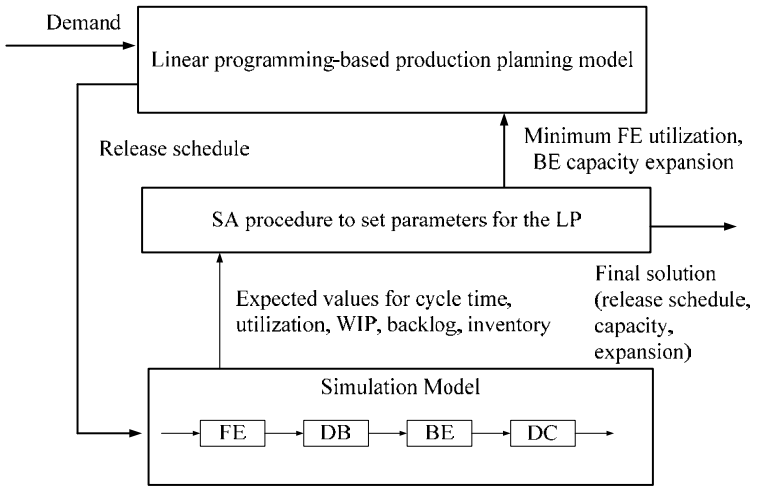


Figure 1: Overall simulation-based optimization scheme.

We use the grid $G := \{(0.5+0.05r, 0.05s) \mid r = 0,\ldots,9, s = 0,\ldots,10\}$ to select arguments $(M,a)$. Of course, we have $|G| = 110$. Therefore, computing the maximum values of $f$ for all the grid points by a full grid search (GS) is time-consuming because of the large number of simulation runs. As a consequence, we are

interested in expediting the search by using a simulating annealing (SA) procedure. For a given solution $(M,a) \in G$ we define the following neighborhood:

$$N(M,a) := \{(M + c_1, a + c_2) \mid c_1, c_2 \in \{-0.05, 0.0, 0.05\}, c_1 + c_2 > 0, (M + c_1, a + c_2) \in G\}, \quad (16)$$

i.e., the neighborhood $N(M,a)$ contains the grid points that are adjacent to the center point. The SA procedure escapes local minima by selecting the current solution $(M',a')$ in a randomized manner. This solution is accepted with probability

$$P((M',a') \mid (M,a)) := \begin{cases} \exp(-\Delta/T), & \text{if } \Delta > 0 \\ 1, & \text{otherwise} \end{cases}. \quad (17)$$

Here, we set $\Delta := f(M,a) - f(M',a')$, where $(M,a)$ is the incumbent solution, i.e. the solution with the largest objective function value found so far. Moreover, the quantity $T$ is called temperature. The SA procedure is based on a geometric cooling scheme for the temperate, i.e., we update the temperature by $T := 0.5T$ . We use the original configuration $(M,a) := (0,0)$ to determine the allowed maximum cycle time values $C_{gt}^*$ by executing the release schedules in the simulation. Moreover, we use $T := f(0,0)$ as initial temperature. The SA procedure terminates when five temperature values in a row do not result in any accepted move, and 16 iterations are performed for each temperature based on recommendations in the literature and some preliminary computational experiments. The search is started from the grid point $(M_0, 0.1)$ where $M_0$ is selected in such a way that the mean utilization value over the horizon obtained from the original configuration is increased by 0.05.

## 4 COMPUTATIONAL EXPERIMENTS

### 4.1 SC Simulation Model

The MIMAC I simulation model is used to represent the FE facility (cf. MIMAC I 2016) while the BE facility is represented by the back-end simulation model (cf. Back-end 2016) proposed by Ehm *et al.* (2011). Note that this model is compatible with the MIMAC I model with respect to size and offered capacity. The SC simulation model contains semiconductor manufacturing characteristics such as batch processing machines, i.e., several lots can be processed at the same time on a single machine, sequence-dependent setup times, exponentially distributed machine breakdowns, operators, and secondary resources. The FE model contains over 200 machines that are organized in 69 work centers. The planned FE bottleneck is given by the stepper work center. First-In-First-Out (FIFO) dispatching is used for the FE facility. The BE model consists of 23 work centers. The FIFO and the Same Setup dispatching rules are used in the BE model. The DB, the DC, and the transportation from the FE facility to the DB and from the BE facility to the DC are represented by machines with an infinite capacity. Two products are considered. FE lots have 48 wafers, whereas BE lots contain only 16 wafers. This means that we have $\Lambda_g = 3$ for $g = 1,2$ . The routes of the first product have 211 and 25 process steps in the FE and the BE facility, respectively, while 246 FE and 31 BE process steps are given for the second product. Transportation activities from the facilities to the DB and DC are represented by additional process steps. Instantaneous material transfer between successive process steps is assumed.

The simulation model is implemented in AutoSched AP. Some customization of the regular Auto-Sched AP framework using the C++ programming language is required to model specifics of the semi-conductor SC. The capacity expansion functionality is provided by activating additional machines using preventive maintenance orders. The resulting simulation model is validated by a domain expert from industry.

### 4.2 Design of Experiments

We expect that the performance of the simulation-based optimization scheme depends on the mean FE bottleneck utilization. Therefore, low and high mean bottleneck utilization levels over a horizon of $T = 15$ periods with weekly periods are considered. A product mix of 1:1 is considered. The planning horizon is divided into three-week subintervals where the utilization is 60% or 80% and 85% or 95% to obtain an average low and high utilization level of 70% and 90%, respectively. Different demand variability values are introduced by varying the coefficient of variation (CV). The demand $D_{gt}$ for product $g$ in period $t$ is generated according to

$$D_{gt} := \Delta_{gt}(1 + z_t), \text{ for all } g \in G, t = 1,\ldots,T, \tag{18}$$

where $\Delta_{gt}$ denotes the mean demand for product $g$ in period $t$ and $z_t$ is a realization of the normally distributed random variable $Z \sim N(0,\sigma^2)$ with $\sigma = CV$. Five independent demand scenarios with low and high mean utilization levels, respectively, are generated. The minimum utilization of the FE machines and the capacity expansion level of the BE facility are chosen from the grid $G$ already specified in Subsection 3.2. 20 independent simulation replications are performed for each single grid point to obtain statistically significant results. The design of experiments is summarized in Table 1.

Table 1: Design of experiments.

| Factor | Level | Count |
|---|---|---|
| Approach | GS, SA | 2 |
| Mean utilization | low, high | 2 |
| CV | 0.10, 0.25 | 2 |
| Minimum utilization of the FE bottleneck | $0.5 + 0.05r$ for $r = 0,\ldots,9$ | 10 |
| Additional capacity of the BE facility | $0.05s$ for $s = 0,\ldots,10$ | 11 |
| Demand scenarios | | 5 |
| Independent simulation replications per factor combination | | 20 |

Both the SA and the GS approach are carried out. The GS requires $(110+1)\cdot 20$ simulation runs per factor combination and the additional reference solution for $(M,a) = (0,0)$. This leads to a total amount of 44400 simulation runs to assess the performance of the GS approach. We will show that the SA approach requires only a fraction of this simulation burden. The realized objective function (15) is only evaluated for the first twelve periods of the planning horizon to avoid end of horizon effects. Initial WIP distributions for each demand scenario are determined by long simulation runs. We use $h_{gt}^{DB} = \omega_{gt}^{F} = b_{gt}^{DC} = 15$ and $h_{gt}^{DC} = \omega_{gt}^{B} = 5$ throughout all experiments. Moreover, we have $r_{gt} = 180$, $\beta = 15$, $\delta = 1000$, and $\lambda = 40$.

We are interested in assessing the degree of constraint violation obtained for executed release schedules. Therefore, we compute the average minimum FE utilization shortfall quantity $ASF := \sum_{t=1}^{12}\left(M\tilde{C}_b - U_{bt}\right)^+ / M\tilde{C}_b$ for each release schedule. The amount of cycle time violation $VCT := \left(C_{gt} - C_{gt}^*\right)^+$ of the realized average cycle time $C_{gt}$ from the maximum allowed cycle time $C_{gt}^*$ for product $g$ in period $t$ averaged over all periods is also of interest. We also report the percentage improvement *Imp* in the objective value (15) of the solution found by the SA approach over the reference solution, i.e., the solution obtained for $(M,a) = (0,0)$. The computing time for a single instance of the

planning model (1)-(12) is on average three minutes on a computer with 3.6 GHz Intel Core™ i7-4790 CPU and 16GB RAM.

### 4.3 Computational Results

First, the solution quality obtained by the SA procedure is investigated. The performance values are averaged over all demand scenarios. The corresponding computational results are shown in Table 2.

Table 2: Performance measure values of the SA approach.

| **Mean utilization** | **CV** | ***ASF* (in %)** | ***VCT* (in min)** | ***Imp* (in %)** |
|---|---|---|---|---|
| Low | 0.10 | 0.68 | 383.10 | 22.22 |
| | 0.25 | 0.99 | 310.77 | 18.35 |
| High | 0.10 | 0.19 | 284.00 | 37.17 |
| | 0.25 | 0.50 | 159.78 | 28.19 |

The cost settings and the chosen penalty and revenue values affect the magnitude of the improvements. Even the executed release schedules obtained by the best performing grid points show minor violations of the maximum allowed cycle time and a slight shortfall in the minimum FE bottleneck utilization due to the variability in the base system. However, significant improvements of the objective function values can be observed. As expected, the amount of improvement is larger for high mean bottleneck utilization since the room for improvement is larger in this situation. Overall, enforcing a minimum utilization of the FE bottleneck work center while expanding the available BE capacities is favorable.

Next, we first show that the SA approach provides release schedules of similar quality compared to those found by the GS scheme. GS determines the parameter combination $(M,a)\in G$ with the highest $f(M,a)$ value by carrying out an exhausted search. The ratio of the realized objective function value obtained from the SA procedure and those found by the GS approach is denoted by SA/GS. These values are shown in Table 3. In addition, the optimal $(M,a)$ parameter combination for each demand scenario found by the SA approach is reported in Table 3.

Table 3: Performance of the SA approach relative to the GS approach.

| **Factor** | **CV** | **Mean utilization** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Low** | | | | | **High** | | | | |
| | | **Demand scenario** | | | | | | | | | |
| | | **1** | **2** | **3** | **4** | **5** | **1** | **2** | **3** | **4** | **5** |
| **SA/GS** | 0.10 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 0.99 |
| $M$ | | 0.70 | 0.80 | 0.80 | 0.75 | 0.80 | 0.90 | 0.90 | 0.85 | 0.90 | 0.50 |
| $a$ | | 0.10 | 0.05 | 0.00 | 0.00 | 0.00 | 0.15 | 0.30 | 0.00 | 0.15 | 0.30 |
| **SA/GS** | 0.25 | 0.98 | 0.76 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 | 1.00 | 0.93 |
| $M$ | | 0.80 | 0.80 | 0.80 | 0.80 | 0.65 | 0.75 | 0.90 | 0.90 | 0.85 | 0.60 |
| $a$ | | 0.15 | 0.10 | 0.05 | 0.10 | 0.25 | 0.00 | 0.05 | 0.30 | 0.10 | 0.50 |

Except for one outlier, the SA approach is able to find optimal or near optimal solutions. The average number of investigated grid points using the SA approach is 19. This leads to a total amount of 7920 simulation runs for the SA approach in contrast to the GS approach that requires 44400 simulation runs.

The best parameter combinations $(M,a) \in G$ with respect to the objective function (15) for a low mean utilization level are obtained for $M$ values that are around ten percent larger than the original planned bottleneck utilization and for slightly expanded capacities in the BE facility. A moderate increase of $M$ compared to the planned bottleneck utilization and additional BE capacity leads to improved profit values as can be seen from Table 2. This effect is larger for low demand variability ($CV = 0.1$). It is interesting to see that for a high mean utilization often $M$ values are selected that are below the planned bottleneck utilization. This avoids a minimum FE utilization shortfall (see Table 2) by smoothing the bottleneck utilization. At the same time, the bottleneck capacity is expanded to avoid maximum cycle time violations (see Table 2) and to improve the profit values.

Overall, the SA procedure is useful for reducing the simulation burden while providing high-quality solutions at the same time. The parameter combination pattern for optimal solutions strongly depends on the planned bottleneck utilization.

## 5 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this paper, we discussed an integrated production planning and capacity expansion problem. A maximum cycle time constraint is assumed for the lots in a simple semiconductor SC. A SA-based simulation-optimization procedure was proposed. Each move of the SA scheme requires solving a linear program to make production planning decisions. The SA scheme selects the minimum bottleneck utilization in the FE facility and the amount of expanded capacity in the BE facility. Increasing the minimum bottleneck utilization in the FE facility might lead to higher throughput in the overall SC and higher average cycle time of the lots in the FE facility while an increasing capacity offered in the BE facility will reduce the cycle time in the SC. Simulation is used to assess the production planning decisions associated with a single move. Simulation experiments using a simplified model of a semiconductor SC including a large-scale FE and a BE model, respectively, were conducted. It turned out that the proposed simulation-optimization scheme is able to outperform conventional planning formulations where capacity expansions in the BE are not taken with respect to profit while a maximum possible cycle time is maintained.

There are several directions for future research. First of all it seems possible to extend the proposed approach to more general SCs, i.e. a network of several wafer fabs, DBs, BE facilities, and DCs should be considered. In addition, more general demand schemes have to be taken into account. We also believe that it is worthwhile to test the presented approach in a rolling horizon setting. Here, the simulation infrastructure proposed by Ponsignon and Mönch (2014) and Ziarnetzky *et al.* (2015) has to be extended to allow for simulation-based decision-making.

## ACKNOWLEDGMENTS

The authors would like to thank Hans Ehm and Thomas Ponsigon, Infineon Technologies AG who pointed out the problem discussed in the present paper. The authors also thank Baris Kacar for providing the clearing function parameters for the work centers of the MIMAC I model.

## REFERENCES

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

Back-end. 2016. “Backend Model.” Accessed April 28, 2016. http://p2schedgen.fernuni-hagen.de/index.php?id=242.

Barahona, F., S. Bermon, O. Günlük, and S. Hood. 2005. “Robust Capacity Planning in Semiconductor Manufacturing.” *Naval Research Logistics* 52(5):459-468.

Bard, J. F., K. Srinivasan, and D. Tirupati. 1999. "An Optimization Approach to Capacity Expansion in Semiconductor Manufacturing Facilities." *International Journal of Production Research* 37(15): 3359-3382.

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

Denton, B., J. Forrest, and R. Milne. 2006. "IBM Solves a Mixed-integer Program to Optimize its Semiconductor Supply Chain." *Interfaces* 36(5):386–399.

Ehm, H., H. Wenke, L. Mönch, T. Ponsignon, and L. Forstner. 2011. "Towards a Supply Chain Simulation Reference Model for the Semiconductor Industry." In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, 2119-2130, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gansterer, M., C. Almeder, and R. F. Hartl. 2014. "Simulation-based Optimization Methods for Setting Production Planning Parameters." *International Journal of Production Economics* 151:206–213.

Hopp, W. J., M. L. Spearman, S. Chayet, K. L. Donohue, and E. S. Gel. 2002. "Using an Optimized Queueing Network Model to Support Wafer Fab Design." *IIE Transactions* 34(2):119-130.

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

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

Liu, J. C. Li, F. Yang, H. Wan, and R. Uzsoy. 2011. "Production Planning for Semiconductor Manufacturing via Simulation Optimization." In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, 3617-3627, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lowe, J., and S. J. Mason. 2016. "Integrated Semiconductor Supply Chain Production Planning." *IEEE Transactions on Semiconductor Manufacturing* 29(2):116-126.

MIMAC I. 2016. "MIMAC Datasets." Accessed April 15, 2016. http://p2schedgen.fernuni-hagen.de/index.php?id=242.

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.

Ponsignon, T., and L. Mönch. 2012. "Heuristic Approaches for Master Planning in Semiconductor Manufacturing." *Computers & Operations Research* 39(3):479–491.

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

Sohn, S. 2004. "Modeling and Analysis of Production and Capacity Planning Considering Profits, Throughput, Cycle Times, and Investment." Ph.D. thesis, H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia. https://smartech.gatech.edu/bitstream/handle/1853/5083/sohn_sugje_200407_phd.pdf [Accessed April 20, 2016].

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.

## 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, production control, and simulation-based optimization. He can be reached by email at Timm.Ziarnetzky@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.