Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

A MULTISTAGE MATHEMATICAL PROGRAMMING BASED SCHEDULING APPROACH FOR THE PHOTOLITHOGRAPHY AREA IN SEMICONDUCTOR MANUFACTURING

Andreas Klemmt
Jan Lange
Gerald Weigert

Electronics Packaging Laboratory
Helmholtzstraße 18
Technische Universität Dresden
01062 Dresden, GERMANY

Frank Lehmann
Jens Seyfert

Infineon Technologies
Königsbrücker Straße 180

01099 Dresden, GERMANY

ABSTRACT

Facilities for wafer fabrication are one of the most complex manufacturing systems. Typically, the bottleneck of such facilities is the photolithography area because of its highly expensive tools and complex resource constraints. In this research, a multistage mixed integer programming based optimization approach for planning of such an area is presented. Thereby, several existing process constraints like equipment dedications, resist allocation, vertical dedications, mask availability are taken into account on the basis of different granularity levels. Altogether eleven different optimization models are presented within four different decomposition stages. Thereby, objected goals are the maximization of throughput, the minimization of setup costs and a balancing of machine utilization. On the basis of real manufacturing data the benefit of the proposed approach is evaluated within a first prototype.

1 INTRODUCTION AND PROBLEM DESCRIPTION

The planning and optimization of semiconductor manufacturing is a very complex task. Especially in the field of wafer processing - the so-called front-end - a lot of different processing steps are performed. These steps are for example typical batch tool operations like oven- and wet-etch processes, or typical cluster tool operations like dry-etch, implant or lithography processes. They have to be repeated to subsequently structure different layers of integrated circuits on the wafers (cf. Figure 1, left). Because of several workcenter-specific constraints and dependencies, these steps are hard to schedule. This is even more complex for facilities with concurrent business modes like production in parallel to research and development processes. That means a wider product mix and a potentially increased number of high-priority lots. As a consequence of complexity the overall scheduling problem is dissected. Also, workcenter-specific optimization approaches are developed.

Usually the photolithography area is a bottleneck workcenter of a wafer fab because of its highly expensive machines and its complex process constraints (cf. Chung and Huang 2008). So, an effective planning of the photolithography area will have a high practical relevance for the whole fab. Generally, in this process a resist is structured to act as a direct mask for subsequent structuring of the underlying substrate material. The photolithography process comprises several sub-processes. Firstly, adhesives are added and moisture is removed from the surface. This is followed by a resist coating, the exposure process and the development of the resist. Finally, there is a curing and an inspection of the resist. The main photolithography process – the exposure – is depicted in Figure 1 (right). Thereby, a reticle (mask) is used to structure a resist layer with the desired circuit pattern. So, for every new layer with a changing pattern, the reticle has to be exchanged. Since integrated circuits are commonly created layer by layer, many cycles of

978-1-4244-9864-2/10/$26.00 ©2010 IEEE

these photolithography processes are performed. Taking into account that channel widths are more and more shrinking, the layer-to-layer alignment, called vertical dedication in the following, is increasingly important for achieving a respectable yield. Even equal lenses are individually slightly different in their characteristics. For this, some wafer always has to be processed on the same lithography unit. Another aspect which adds to the complexity of lithography is the application of different resists. There are specific resists and specific resist thicknesses needed for each operation. Whenever a different kind of resist thickness is applied, an inspection of the layer characteristics has to be carried out.

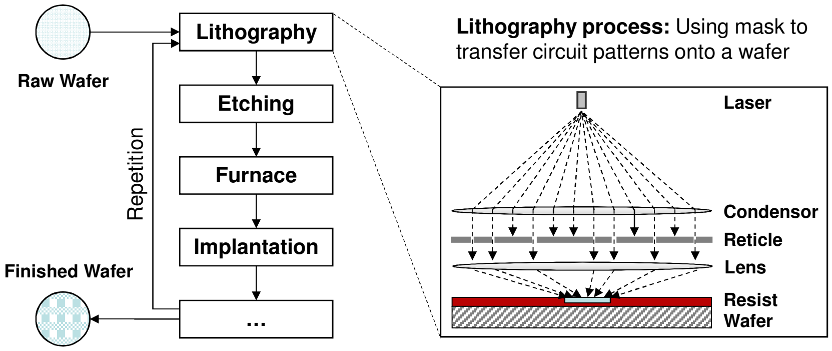


Figure 1: The photolithography process (cf. Chung and Huang 2008)

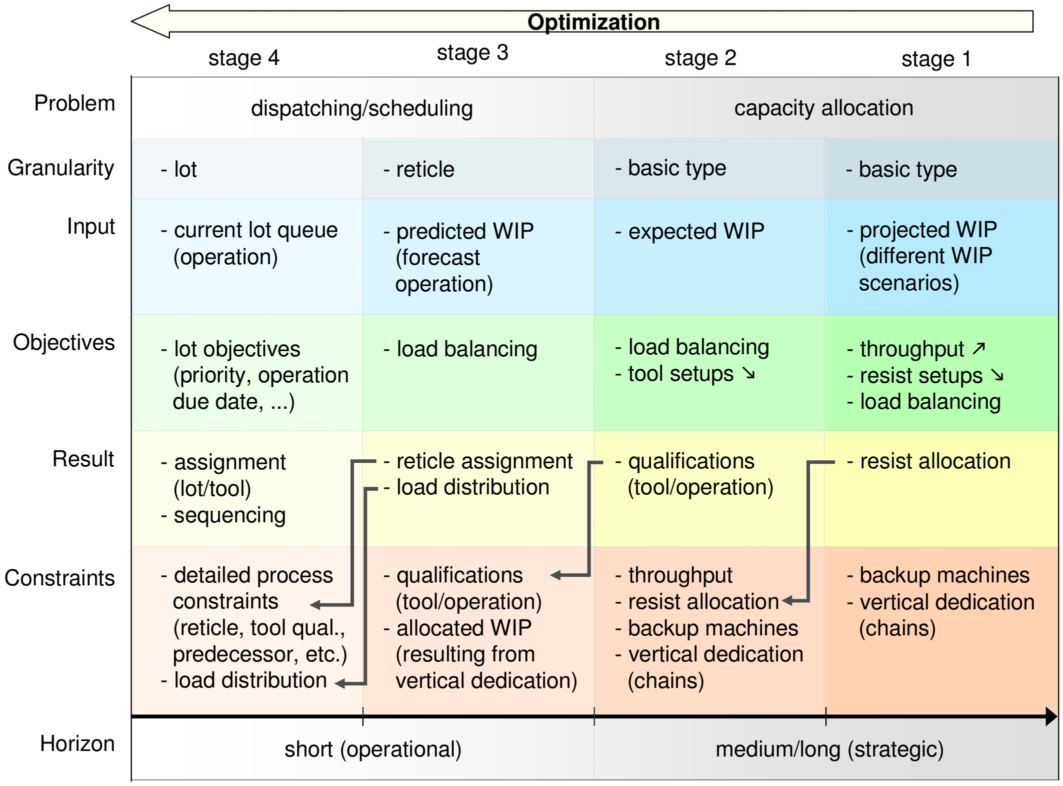
So, a pure lot-based scheduling approach is not sufficient for optimizing a lithography process. Moreover, also tool-setups and resist-setups have to be taken into account to react on the changing product-mix. Therefore, this paper presents a multi-stage mathematical programming based approach for planning and scheduling in the lithography area which focuses on different objectives and granularity levels.

2 PREVIOUS RELATED WORK

Approaches for optimizing lithography processes are a matter of particular interest and are also recorded in several publications. Toktay and Uzsoy (1998) developed a shift-based mixed integer programming model for short time capacity allocation. Capacity allocation does not mean to find detailed production schedule but to assign a certain amount of operation WIP (Work-In-Process) to a machine. An operation thereby is a combination of layer and product. The investigated performance measure of this approach was throughput maximization. Toktay and Uzsoy (1998) show that the capacity allocation problem with integer side constraints (maximum number of machines per operation, maximum number of operations per machine) is strongly NP-hard. This problem was also discussed by Akcali, Üngör and Uzsoy (2005). They transformed the capacity allocation problem into a network flow problem. Thereby several approximation heuristics are discussed. A much more detailed MIP model for short time capacity allocation is given by Kim, Yea and Kim (2002). Their objective was to find a machine allocation meeting target production quantities as well as possible. However, the approaches above do not consider all existing lithography process constraints even from a pure capacitive viewpoint. They neglect the vertical dedication restriction (layer to layer assignment) which is essential to good quality of final products (cf. Chung, Huang and Lee 2006). A MIP model primary focussing on this constraint is given by Pham, Shr and Chen (2008). Here the optimization objective is to find an order-layer-tool assignment minimizing a production cost function. Much more detailed MIP models regarding the layer-to-layer constraint are given by Chung, Huang and Lee (2006) and Chung, Huang and Lee (2008). Here also the optimization objective is to find a order-layer-tool assignment. Furthermore a load balancing within time buckets is discussed. A load balancing approach performed on the basis of a detailed simulation model is discussed in Mönch, Prause and Schmalfuss (2001). An other simulation-based approach focusing on dispatching policies can be found in Akcali, Nemoto and Uzsoy (2001). In this work it is primarily investigated the objective cycle time. An approach combining simulation and artificial intelligence is presented in Arisha, A. and P. Young (2004). Here, the primary goal was the minimization of WIP, setup time and throughput time.

3 SOLUTION APPROACH

In this section it is presented a multistage mixed integer programming based decomposition approach for optimizing the manufacturing flows in lithography area. The motivation of this decomposition is simply due to the fact that the overall problem (with all its restrictions an the number of inputs) is too complex for efficiently handle it in one mathematical model. Furthermore, the arising problem viewpoints are very different and depend on the problem horizon (strategic capacity planning vs. operational planning/scheduling). In literature mostly one kind of problem was picked out and investigated, neglecting other influenceable parameters. So, optimizing the lot sequencing/scheduling decision on shop floor level will not lead to significant benefits if the tool qualifications are not optimal for the current WIP scenario. These qualifications are also not only 0/1 decisions (e.g. process possible, process not possible). They sometimes depend on additional resources (e.g. process after resist installation possible). Because of that, in this research the overall scheduling problem is dissected in several optimization stages (cf. Figure 2). Thereby, in every stage different objectives are optimized with regard to different kinds of process constraints, data granularity and planning horizons. Every stage can be optimized/executed as stand alone (classical, fixed input) or as a combination of overlying optimization stages (relaxed input). In this case every optimization stage reduces the degree of freedom for the next stage leading finally to the decision making on the shop floor level. These four stages will be discussed more detailed in the next sections.



Optimization

	stage 4	stage 3	stage 2	stage 1
Problem	dispatching/scheduling		capacity allocation	
Granularity	- lot	- reticle	- basic type	- basic type
Input	- current lot queue (operation)	- predicted WIP (forecast operation)	- expected WIP	- projected WIP (different WIP scenarios)
Objectives	- lot objectives (priority, operation due date, ...)	- load balancing	- load balancing - tool setups ↘	- throughput ↗ - resist setups ↘ - load balancing
Result	- assignment (lot/tool) - sequencing	- reticle assignment - load distribution	- qualifications (tool/operation)	- resist allocation
Constraints	- detailed process constraints (reticle, tool qual., predecessor, etc.) - load distribution	- qualifications (tool/operation) - allocated WIP (resulting from vertical dedication)	- throughput - resist allocation - backup machines - vertical dedication (chains)	- backup machines - vertical dedication (chains)
Horizon	short (operational)		medium/long (strategic)	

Figure 2: Problem decomposition

3.1 Stage 1: Long time capacity allocation – resist installation

The goal of optimization stage 1 is to find a tool qualification (inclusive resists installations/changes) which is robust against different possible WIP scenarios projected for the next weeks and which coevally allows a stable and balanced tool utilization. Because of the projected WIP's uncertainty, the planning granularity is reduced to the basic type level (collection of related products with identical process parameters but different masks). That means all lots of one basic type are cumulated. In the following, the combinations of this cumulated WIP per basic type and its different layers are denoted as WIP per operation. On this level of abstraction, lot based constraints and objectives are completely neglected or only modelled capacity based (e.g. vertical dedications as chains). To further react on uncertainties, the allocation of backup machines (with all necessary process constraints) for critical operations is a goal of this stage. The following notations are defined as follows (some of them are also used in later sections):

The lithography workcenter consists of m (partly different) machines M_k (k =1,…m). There are n different operations O_i (i =1,…n) assigned to the lithography process. A dedication matrix $A := (A_{ik})^{nxm} \in \{0,1\}$ specifies if operation O_i is basically technological feasible (e.g. laser, wavelength) for processing on machine M_k with a processing time $p_{ik} > 0$. For every layer O_i exists a predefined WIP $W_i > 0$ in the planning horizon. Also for every layer O_i a minimal number $M_i^{\min}$ and maximal number $M_i^{\max}$ of machines and a number M_i^{back} of backup machines to be qualified is defined. Moreover, for every machine M_k a predefined machine availability $V_k > 0$ exists in the planning horizon to regard possible planned tool down times d_k. Also, for every machine M_k it is defined a maximal number of $O_k^{\max}$ operations to be qualified. A set D_e includes all operations belonging to the same vertical dedication (e =1,…,q). Thereby, q is the number of all vertical dedications. Furthermore, there are r different resists R_l (l =1,…,r) and t different resist-thicknesses-combinations T_s (s =1,…t) with ($t < r$). A matrix $T^O := (\ T_{is}^{\ O}\ )^{nxt} \in \{0,1\}$ specifies if operation O_i requires the resist-thicknesses-combinations T_s. Another matrix $T^M := (\ T_{sk}^{\ M}\ )^{txm} \in \{0,1\}$ specifies if machine M_k is currently equipped for the resist-thicknesses-combinations T_s. Furthermore, there a surjective function f mapping each resist-thicknesses-combination T_s ($s \in \{1,\dots,t\}$) to exactly one resist R_l ($l \in \{1,\dots,r\}$). A matrix $R^O := (R_{il}^{\ O})^{nxr} \in \{0,1\}$ specifies if operation O_i requires resist R_l. A further matrix $R^M := (R_{lk}^{\ M})^{rxm} \in \{0,1\}$ states if resist R_l is currently installed on machine M_k. Note: Matrix R^M and R^O are directly derived from T^M and T^O via f. The following minimal example should illustrate these coherences:

Example: For processing operation O_1 and operation O_2 on machine M_1 the resist R_1 has to be installed on the machine. However, both operations require different resist thicknesses.

$$\begin{array}{c|cc|} T_{is}^{\ O} & T_1 & T_2 \\ \hline O_1 & 1 & 0 \\ O_2 & 0 & 1 \\ \hline \end{array} \qquad \begin{array}{c|c|} T_{sk}^{\ M} & M_1 \\ \hline T_1 & 1 \\ T_2 & 1 \\ \hline \end{array} \qquad \begin{array}{c} \rightarrow \\ f(T_1) = R_1 \\ f(T_2) = R_1 \end{array} \qquad \begin{array}{c|c|} R_{il}^{\ O} & R_1 \\ \hline O_1 & 1 \\ O_2 & 1 \\ \hline \end{array} \qquad \begin{array}{c|c|} R_{lk}^{\ M} & M_1 \\ \hline R_1 & 1 \\ \hline \end{array}$$

Beside the described input parameters some index sets have to be defined. These are helpful for the mathematical model present later:

$A_i := \{k \mid A_{ik} = 1\}$ set of all machines M_k which are able to process operation O_i.

$A_k := \{i \mid A_{ik} = 1\}$ set of all operations O_i which can be processed on machines M_k.

$R_i^O := \{l \mid R_{il}^O = 1\}$ set of all resists R_l which are required for operation O_i.

$T_i^O := \{s \mid T_{is}^O = 1\}$ set of all resist-thicknesses-combinations T_s which are required for operation O_i.

$O^{back} := \{i \mid M_i^{back} \geq 1\}$ set of all operations O_i which can be processed on machines M_k.

By the help of the index sets it is possible to encode only information in the solution vector which are really essential for problem description. Now the following unknowns (the solution vector) of the mathematical model are defined:

$x_{ik} \in \mathbb{R}_+$ amount of wafers in operation O_i assigned to machine M_k; (k =1,…,m; $i \in A_k$).

$y_{ik} \in \{0,1\}$ operation O_i is qualified on machine M_k, 0 otherwise; (k =1,…,m; $i \in A_k$).

$z_{ik} \in \{0,1\}$ operation O_i is qualified as backup on machine M_k, 0 otherwise; ($i \in O^{\text{back}}$; $k \in A_i$).

$v_{lk} \in \{0,1\}$ resist R_l is installed in machine M_k, 0 otherwise; (k =1,…,m; l =1,…,r).

$w_{sk} \in \{0,1\}$ combination T_s is installed in machine M_k, 0 otherwise; (k =1,…,m; s =1,…t).

$b^U \in \mathbb{R}_+$ upper bound for machine utilization.

$b^L \in \mathbb{R}_+$ lower bound for machine utilization.

$b_k^U \in \mathbb{R}_+$ upper bound for utilization of machine M_k; (k =1,…,m).

$b^{U_R} \in \mathbb{R}_+$ upper bound for resist new-installation.

$b^{L_R} \in \mathbb{R}_+$ lower bound for resist de-installation.

$b^{U_T} \in \mathbb{R}_+$ upper bound for resist-thicknesses-combinations.

$P_{\max} \in \mathbb{R}_+$ throughput.

The optimization process of stage 1 is now performed in four steps (optimization model 1-4). In the first model the objective is the maximization of throughput. This optimization is primarily a test if the projected WIP is achievable with regard to the basic process constraints mentioned above (machine availability, max/min machines per operation, etc.).

Optimization model 1 (throughput maximization):

$$P_{\max} \to \max \qquad \text{subject to} \tag{1}$$

$$\sum_{k \in A_i} x_{ik} \le W_i \qquad i = 1,...,n \tag{2}$$

$$M_i^{\min} \le \sum_{k \in A_i} y_{ik} \le M_i^{\max} \qquad i = 1,...,n \tag{3}$$

$$\sum_{i \in A_k} p_{ik} x_{ik} \le V_k \qquad k = 1,...,m \tag{4}$$

$$\sum_{i \in A_k} y_{ik} \le O_k^{\max} \qquad k = 1,...,m \tag{5}$$

$$\sum_{k \in A_i} z_{ik} \quad = M_i^{back} \qquad i \in O^{\text{back}} \tag{6}$$

$$y_{ik} + \alpha \le v_{lk} \qquad i = 1,...,n;\ k \in A_i;\ l \in R_i^O;\ \alpha = \{z_{ik} \text{ if } i \in O^{\text{back}}; 0 \text{ else}\} \tag{7}$$

$$y_{ik} + \alpha \le w_{sk} \qquad i = 1,...,n;\ k \in A_i;\ s \in T_i^O;\ \alpha = \{z_{ik} \text{ if } i \in O^{\text{back}}; 0 \text{ else}\} \tag{8}$$

$$x_{ik} = x_{jk} \qquad e = 1,...,q;\ i, j \in D_e;\ i \ne j;\ k \in A_i = A_j \tag{9}$$

$$y_{ik} - \frac{x_{ik}}{PW_i} \le 0 \qquad i = 1,...,n;\ k \in A_i \tag{10}$$

$$\min\left(PW_i, \frac{V_k}{p_{ik}}\right) y_{ik} \ge x_{ik} \qquad i = 1,...,n;\ k \in A_i \tag{11}$$

$$\sum_{i=1}^{n} \sum_{k \in A_i} x_{ik} \ge P_{\max} \tag{12}$$

Thereby equation (12) restricts objective function (1). Equation (2) restricts the assigned WIP to the machines to available WIP. Constraint (3) ensures the minimum/maximum number of machine qualifications. Equation (4) limits the cumulated machine work to their availability boundaries. Constraint (5) restricts the number of qualified operations per machine. Equation (6) forces that the number of backup machines per operation is exactly considered. Constraints (7) and (8) ensure that for every qualified (or backup qualified) operation the resists and the resists thickness combination is installed. Equation (9) implements the vertical dedication constraint. Thereby, i is the first element of D_e. Constraint (10) ensures that if a machine is qualified for a specific operation at least the P-th part (P = input parameter) of the operations WIP is assigned to the machine. Otherwise, equation (11) assures that if a specific WIP per operation is assigned to a machine, this machine has to be qualified for the operation.

In the following steps of stage 1 the maximal throughout $P_{\max}$ is fixed by constraint (12) to its optimal value calculated by optimization model 1. So only solutions are further regarded allowing this maximal throughput. In the next optimization step an upper bound for workload balancing is calculated. This results form the following mathematical model:

Optimization model 2 (load balancing – upper bound):

$$b^U \rightarrow \min \qquad \text{subject to} \tag{13}$$

$$\sum_{k=1}^{m} \sum_{i \in A_k} p_{ik} x_{ik} \le b^U \tag{14}$$

and constraints (2) - (12).

In this model constraint (14) restricts the objective function (13). That means the maximal workload of the most utilized machine is decreased as much as possible. All other constrains (2) - (12) are untouched and are furthermore active. In all following steps of long time capacity allocation this minimal upper bound is fixed by constraint (14) to its optimal value calculated by optimization model 2. So, further only solutions are regarded allowing the maximal throughput and the minimized maximal utilization. In the next optimization step the installation/setup of resists is minimized. Also, the number of resist thickness combinations on the machines is minimized to reduce the number of future inspections. The following mathematical program is solved:

Optimization model 3 (resists optimization):

$$\omega_1 b^{U_R} - \omega_2 b^{L_R} + \omega_3 b^{U_T} \rightarrow \min \qquad \text{subject to} \tag{15}$$

$$\sum_{k=1}^{m} \sum_{l=1 (R_{lk}^M = 0)}^{r} v_{lk} \le b^{U_R} \tag{16}$$

$$b^{L_R} + \sum_{k=1}^{m} \sum_{l=1 (R_{lk}^M = 1)}^{r} v_{lk} \le \sum_{k=1}^{m} \sum_{l=1}^{r} R_{lk}^M \tag{17}$$

$$\sum_{k=1}^{m} \sum_{s=1}^{t} w_{sk} \le b^{U_T} \tag{18}$$

and constraints (2) - (12), (14).

In this model a multi-criteria objective function (15) is optimized. This function is restricted by the three constraints (16), (17) and (18). Thereby, equation (16) calculates an upper bound for resist-new-installations. Equation (17) calculates a lower bound for resist-de-installations and equation (18) restricts the total number of resists-thickness-combinations. By the help of the weighting parameters ω_1, ω_2 and ω_3 a prioritization of the three combined objectives: minimization of resist-new-installations, maximization of resist-de-installations and minimization of resists-thickness-combinations is possible. Typically, it is chosen $\omega_1 >> \omega_2$ and $\omega_1 >> \omega_3$. Now, in the last optimization step all resist bounds (results from optimization model 3 added by potentially offsets) are fixed by the constraints (16), (17) and (18). It follows the optimization of the load balancing lower bounds by the following mathematical model:

Optimization model 4 (load balancing – lower bound):

$$\omega_4 b_k^U - \omega_5 b^L \to \min \qquad \text{subject to} \tag{19}$$

$$\sum_{k=1}^{m}\left(\sum_{i\in A_k} p_{ik}x_{ik} + d_k\right) \ge b^L \tag{20}$$

$$\sum_{i\in A_k} p_{ik}x_{ik} \le b_k^U \qquad k = 1,\dots,m \tag{21}$$

and constraints (2) - (12), (14), (16) - (18).

In this model constraint (20) restricts the objective function (19). That means the minimal workload of the least utilized machine is increased as much as possible ($\omega_5 >> \omega_4$). By the help of equation (21) the workload of every machine is additionally decreased. This has the effect that fast machines are preferably used. Note: Optimization model 2, 3 and 4 are mutually dependent. So, hard bounds for resist installation can lead to low quality balancing solutions. Because of that, it is essential to define some offsets concerning these bounds. They can be found by iterating step 2, 3 and 4 several times. Figure 2 illustrates the long time capacity optimization scenario of stage 1.

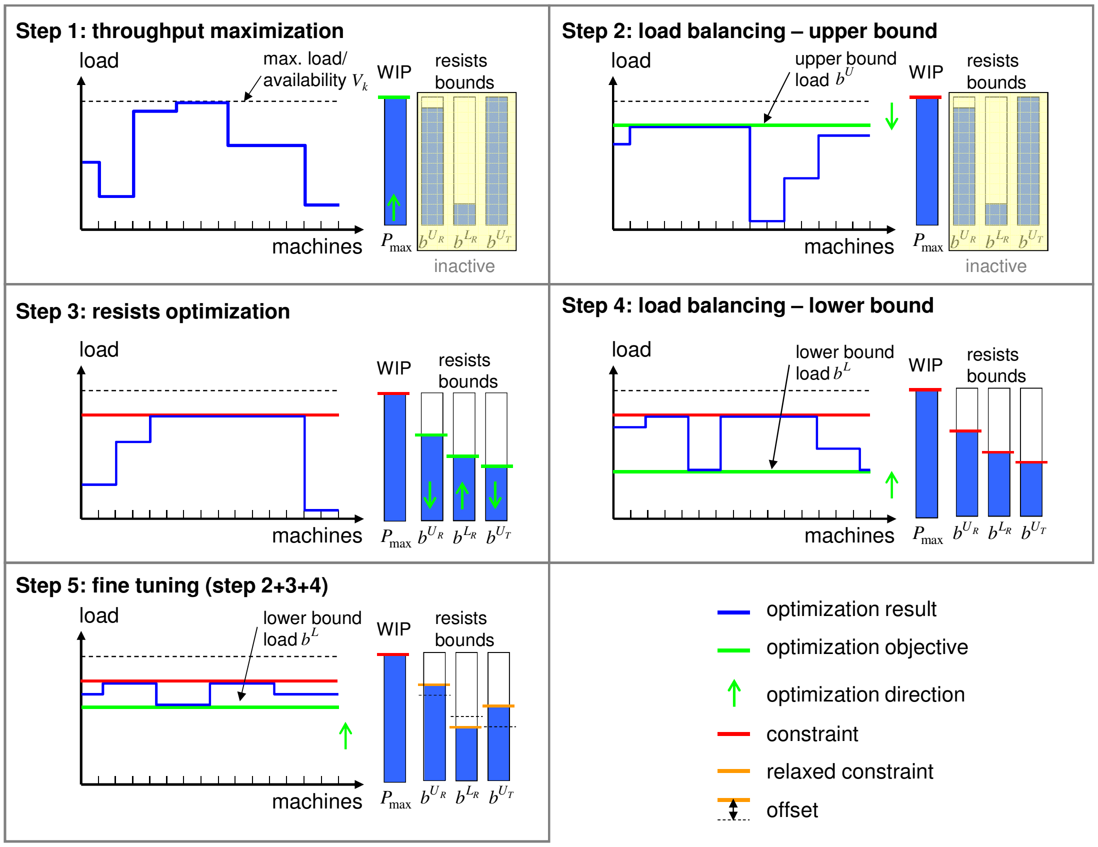


Figure 3: Optimization steps for long time capacity allocation

Because of uncertainties in future WIP trends it is necessary to take into account different projected/possible WIP scenarios. So the long time capacity allocation has to be optimized several times with regard to these different scenarios. Thereby the overall objective is primarily to find a resist's allocation which is robust against mentioned uncertainties.

3.2 Stage 2: Medium time capacity allocation – tool qualification

The result of the long time capacity allocation is an optimized resist allocation which is robust against different possible WIP scenarios and which coevally allows a balanced tool utilization. Every solution of stage 1 also leads to an operation/tool qualification matrix (variable y_{ik}) which is more or less a casual information of this stage. Now, in the medium time capacity allocation this qualification matrix is optimized with respect to the (changing) expected WIP. Therby the goal is to minimize the number of necessary new qualifications. Because of the shorter planning horizons the reputation rate of this stage is higher than in stage 1. Thereby, the resists allocation of stage 1 is fixed. Consequently, the dedication matrix $A := (A_{ik})^{nxm} \in \{0,1\}$ now specifies if operation O_i can be processed on machine M_k (process technological able and resist installed). Further a matrix $I := (I_{ik})^{nxm} \in \{0,1\}$ specifies if operation O_i is currently qualified on machine M_k.

The optimization of stage 2 is now performed in 3 steps. Using the variable definitions and constraints of section 3.1 (all variables and constraints concerning resists are dropped) a first optimization step for medium time capacity allocation calculates an upper bound for workload balancing similar to optimization model 2:

Optimization model 5 (load balancing – upper bound):

$$b^U \to \min \qquad \text{subject to} \tag{13}$$

and constraints (2) - (6), (9) – (12), (14).

Also the optimization of load balancing lower bounds is adapted to this scenario:

Optimization model 6 (load balancing – lower bound):

$$\omega_4 b_k^U - \omega_5 b^L \to \min \qquad \text{subject to} \tag{19}$$

and constraints (2) - (6), (9) – (12), (14), (20) - (21).

The model is also formally identical to optimization model 4 (without arc allocations). Now, in the last optimization step all load balancing bounds (results from optimization model 5 and 6 added by potentially offsets) are fixed by the constraints (14), (20) and (21). It follows the minimization of necessary new qualifications concerning the current operation/machine allocation to reach a balanced solution:

Optimization model 7 (setup minimization – new qualification):

$$\sum_{i=1}^{n} \sum_{k \in A_i (I_{ik}=0)} y_{ik} + \alpha \to \min \qquad \text{subject to} \tag{22}$$

and constraints (2) - (6), (9) – (12), (14), (20) - (21).

Thereby α is defined as α = {z_{ik} if $i \in O^{\text{back}}$; *0* else}. Note that optimization model 6 and 7 are also mutually dependent. So, hard bounds for load balancing can lead to a high number of new qualifications. Slightly relaxing these bounds can lead to solutions where only a couple of new qualifications become necessary. They can be found by iterating step 6 and 7, too.

3.3 Stage 3: Short time capacity allocation – reticle scheduling

The result of the long and medium time capacity allocation is an optimized machine qualification matrix (y_{ik}) which allows a balanced tool utilization. Therefore, every solution of the previous stages implicates an operation/tool allocation matrix (variable x_{ik}). However, this information was not used so far. Rather it was treated as a proof of concept concerning tool qualification from a static viewpoint. Now, in the short time capacity allocation this matrix is calculated for the current shop floor which is the input for the operating system (next stage; cf. section 3.4). The result of this stage is a static reticle assignment for short time horizons. However, on the shop floor level now dynamic aspects have to be regarded, too. So the reputation rate of this stage is very high, to react on changing WIP scenarios. Furthermore, the granularity level has to be increased. That means all lots requiring the same reticle are cumulated. So, in this stage the combinations of the cumulated WIP per reticle are now denoted as WIP per operation. Also, only currently active machines are regarded. So V_k, d_k and $O_k^{\max}$ are dropped. All other input parameters have to be adapted the same way to this granularity level. Furthermore, all tool qualifications (result of stage 2) are now fixed. That means, the dedication matrix $A := (A_{ik})^{nxm} \in \{0,1\}$ specifies if operation O_i is qualified (or qualified as backup) on (available) machine M_k. In contrast to stage 2 W_i now represents the WIP at the operation. Also, $M_i^{\max}$ now specifies the exact number of reticles for operation O_i.

Furthermore, dynamic (lot depending) parameters have to be regarded: The amount of lots having a vertical dedication is considered. That means for some lots of the operations O_i that a predefined minimum capacity allocation N_{ik} exists for a machine M_k. Also, lot priorities have to be observed. For this every lot gets a priority value (e.g. normal lot = 1, engineering lot = 5, rocket lot = 100). Now, a priority w_{ik} is calculated for every operation and qualified machine. This is the sum of the priority values of all lots which are able for processing on M_k (with respect to the vertical dedication constraints). Furthermore, dynamic machine properties have also to be regarded. For this, a cost value c_{ik} is calculated for every operation and qualified machine. This value specifies a relative cost which will arise if the reticle of operation O_i is moved to machine M_k (e.g. reticle in machine = 0, reticle in stocker = 1, send-ahead lot necessary = 5, focus time out = 50, send-ahead lot impending = -5). So, a negative cost implies a benefit for moving the reticle to prevent impending costs. In analogy to section 3.1 the following variables have to be defined for short time capacity allocation:

$x_{ik} \in \mathbb{R}_+$ amount of wafers for reticle of O_i assigned to machine M_k; (k =1,…,m; $i \in A_k$).
$y_{ik} \in \{0,1\}$ reticle of O_i is assigned to machine M_k, 0 otherwise; (k =1,…,m; $i \in A_k$).
$B_{\max} \in \mathbb{R}_+$ cost value.

The optimization process of stage 3 is now performed in four steps (optimization model 8-11). In the first model the objective is the maximization of priority allocations and the minimization of machine costs.

Optimization model 8 (Priority allocation and machine costs):

$$B_{\max} \rightarrow \max \qquad \text{subject to} \tag{23}$$

$$y_{ik} - x_{ik} \leq 0 \qquad i = 1,...,n;\ k \in A_i \tag{24}$$

$$x_{ik} - W_i y_{ik} \leq 0 \qquad i = 1,...,n;\ k \in A_i \tag{25}$$

$$N_{ik} y_{ik} - x_{ik} \leq 0 \qquad i = 1,...,n;\ k \in A_i; N_{ik} > 0 \tag{26}$$

$$N_{ik} + \sum_{l \in A_i\,(l \neq k)} x_{il} \leq W_i \qquad i = 1,...,n;\ k \in A_i; N_{ik} > 0 \tag{27}$$

$$x_{ik} + \sum_{l \in A_i\,(l \neq k)} N_{ik} \leq W_i \qquad i = 1,...,n;\ k \in A_i \tag{28}$$

$$\sum_{i=1}^{n} \sum_{k \in A_i} \left(\frac{w_{ik}}{M_i^{\max}} - c_{ik} \right) y_{ik} \leq B_{\max} \tag{29}$$

and constraints (2), (3).

With the help of the cost value (23) only reticle allocations are considered which are beneficial regarding current lot priorities and machine states. So, by constraint (29) this value will be fixed (possibly relaxed) in all further stages) to its optimal value calculated by optimization model 8. Constraint (24) and (25) are adaptations of (10) and (11). The Equations (26) - (28) implement the vertical dedication scenario on the new granularity level. The Constraints (2) and (3) are formally identical but also adapted. In the next model the throughput is maximized within the restricted/optimized reticle allocation options:

Optimization model 9 (Throughput maximization):

$$P_{\max} \to \max \qquad \text{subject to} \qquad (1)$$

and constraints (2), (3), (12), (24) - (29).

In the following steps of stage 3 constraint (12) fixes the maximal throughout $P_{\max}$ (possibly relaxed) to its optimal value calculated by optimization model 9. Similar to previous stages in the last two models the load balancing boundaries are optimized:

Optimization model 10 (load balancing – upper bound):

$$b^U \to \min \qquad \text{subject to} \qquad (13)$$

and constraints (2), (3), (12), (14), (24) - (29).

Optimization model 11 (load balancing – upper bound):

$$\omega_4 b_k^U - \omega_5 b^L \to \min \qquad \text{subject to} \qquad (19)$$

and constraints (2), (3), (12), (14), (20), (21), (24) - (29).

3.4 Stage 4: Lot sequencing

The result of stage 3 is an optimized reticle and a balanced capacity allocation for a short time horizon. Now, these information are used by the dispatching system for decision making on shop floor level. Here, several other process constraints (which are considered only capacitive so far) like send-ahead wafers/lots, focus time outs, etc. are observed too. The objectives for assigning the lots the machines and for sequencing are primarily lot-based (due dates, lot priorities). However, the inputs generated by the solver-based pre-calculations/optimizations restrict the degree of freedom of the dispatching system significantly (in most cases only one reticle is available for an operation). This has the advantage, that the complexity of the dispatching rules can be reduced significantly.

In future it is also planned to develop mathematical programming based cluster tools models to further optimize the sequencing scenario. Therefore the approaches shown in Klemmt et al. (2009) will be further investigated. Lot sequencing control may consider expected material arrival (look ahead, batch building), due date based lot priorities and quality or sampling related lot-tool assignments in more detail.

4 APPLICATION AND RESULTS

By now all models of stage 1, 2 and 3 are implemented in one software framework (cf. Figure 4). Thereby, the input data is extracted from different source types (MS-Excel sheets, databases, etc.). A graphical user interface (GUI) allows the visualisation and also the manipulation of this data. This is (especially for static capacity planning) a very crucial point. So a system or process expert can also test new scenarios (buying new equipments, installing new operations, new resists, etc.) and directly identify their effects on different granularity levels. Therefore the GUI allows the visualisation or respectively the export of the results. Furthermore, the GUI also provides other options like the parameterization of the weighting, cost and priority parameters mentioned above as well as solver settings (time limits, gaps, etc).

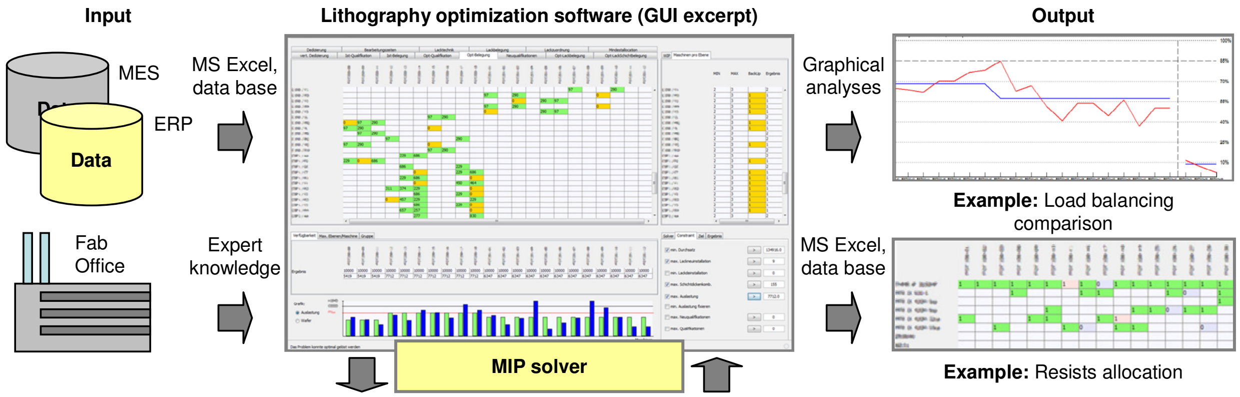

Input
MES
ERP
Data
Fab Office
MS Excel, data base
Expert knowledge
Lithography optimization software (GUI excerpt)
MIP solver
Output
Graphical analyses
Example: Load balancing comparison
MS Excel, data base
Example: Resists allocation

Figure 4: Optimization framework (data obliterated)

The problems to be solved in practical application are high dimensional. That means there are dozens of machines and hundreds of different operations to be planned. However, by keeping the number of unknowns problem adapted (with the help of index sets; see section 3) all problems are mostly solvable within a few seconds.

Without disclosing confidential information it can be stated that the presented approach has two major benefits in the production environment: On the on hand there are notable improvements in load balancing (cf. Figure 4; upper right), resists allocation and the minimization of inspections. On the other hand, the effort in calculating a good/optimized capacity allocation, which was done by hand so far, is drastically decreased. Also a lot of interaction between the different planning stages is possible now. Automated data base updates will further reduce model administration efforts and are a precondition for planning stage 3. Manual reticle disposition on shop floor level will be replaced by fully automated decision making and execution.

5 CONCLUSIONS AND FUTURE RESEARCH

At first, most of the optimization problems, especially in the semiconductor manufacturing, appear too complex for solving by mathematical programming based methods. This was the same in the case of photolithography area and its required objectives. But it could be shown that the problem is solvable not only in an academic sense but also for operative application in the shop floor. The key is the segmentation of the problem into four stages and the creation of suitable data interfaces between them. Of course, the global optimum is possibly lost but the optimization process itself is essentially accelerated. The investigations has discovered a notable optimization potential in comparison to the manual planning methods.

The promising results and positive production feedback we got from lithography so far encourages us to go ahead and apply the multi stage approach to other process areas. Dynamic implant setup optimization or short term tester allocation including probe card assignment are examples of comparable pre-optimization scenarios which could be integrated with lot sequencing in the multi stage model.

ACKNOWLEDGMENTS

This work was supported by Infineon Technologies Dresden, by the EFRE fund of European Union and funding of the State Saxony (project number 13139/2219).

REFERENCES

Akcali, E., K. Nemoto, and R. Uzsoy. 2001. *Short Cycle-Time Improvements for Photolithography Process in Semiconductor Manufacturing*. In *Transactions Semiconductor Manufacturing* 14:48-56.

Akcali, E., A. Üngör, and R. Uzsoy. 2005. *Short-Term Capacity Allocation Problem with Tool and Setup Constraints*. In *Naval Research Logistics* 52:754-764.

Arisha, A. and P. Young. 2004. *Intelligent simulation-based lot scheduling of photolithography toolsets in a wafer fabrication facility*. In *Proceedings of the 36th conference on Winter simulation*, 1935-1942.

Chung, S.H., C.Y. Huang, and A.H.I. Lee. 2008. *Heuristic algorithms to solve the capacity allocation problem in photolithography area (CAPPA)*. In *OR Spectrum* 30:431-452.

Chung, S.H., C.Y. Huang, and A.H.I. Lee. 2006. *Using Constraint Satisfaction Approach to Solve the Capacity Allocation Problem for Photolithography Area*. In *Computational Science and Its Applications* 3982:610-620.

Kim, S., S.-H. Yea, and K. Bokang. 2002. *Shift scheduling for steppers in the semiconductor wafer fabrication process*. In *IIE Transactions* 34:167-177.

Klemmt, A., S. Horn, G. Weigert, G. and K.-J. Wolter. 2009. *Simulation-based optimization vs. mathematical programming: A hybrid approach for optimizing scheduling problems*. In *Robot. Comput.-Integr. Manuf.* 25:917-925.

Mönch, L., M. Prause, and V. Schmalfuss. 2001. *Simulation-based solution of load-balancing problems in the photolithography area of a semiconductor wafer fabrication facility*. In *Proceedings of the 33nd conference on Winter simulation*, 1170-1177.

Pham, H.N.A., S. Shr, and P.P. Chen. 2008. *An Integer Linear Programming Approach for Dedicated Machine Constraint*. In *Seventh IEEE/ACIS International Conference on Computer and Information Science*, 69-74.

Toktay, L.B. and R. Uzsoy. 1998. *A capacity allocation problem with integer side constraints*. In *European Journal of Operational Research* 109:170-182.

AUTHOR BIOGRAPHIES

ANDREAS KLEMMT studied mathematics at Dresden University of Technology, Germany. He obtained his degree in 2005 in the field of optimization. He has been a Research Assistant at Electronics Packaging Laboratory of the Dresden University of Technology since 2006 and works on the field of production control, simulation & optimization of manufacturing processes, especially in the field of electronics and semiconductor industry. His email is <klemmt@avt.et.tu-dresden>.

JAN LANGE received his master's degree in Information Systems Technology at the Dresden University of Technology in 2008. He is now a Research Assistant at the Electronics Packaging Laboratory of the Dresden University of Technology and works in the field of production control, simulation and optimization of manufacturing processes. His email is <lange@avt.et.tu-dresden.de>.

GERALD WEIGERT is an Assistant Professor at Electronics Packaging Laboratory of the Dresden University of Technology. Dr. Weigert works on the field of production control, simulation & optimization of manufacturing processes, especially in electronics and semiconductor industry. His email is <Gerald.Weigert@tu-dresden.de>.

FRANK LEHMANN received his masters degree in Electrical Engineering at the Dresden University of Technology. He works as a staff engineer within the Process and Productivity Mastering group of Infineon Dresden and is responsible for WIP flow management improvement projects. His email address is <frank.lehmann@infineon.com>.

JENS SEYFERT obtained his diploma degree in Microtechnologies at the University of Applied Sciences Zwickau. Since 2000 he has been a manufacturing engineer at the lithography department of Infineon Dresden. His email address is <jens.seyfert@infineon.com>.