Proceedings of the 2011 Winter Simulation Conference
S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds.

AN OPTIMIZATION APPROACH FOR PARALLEL MACHINE PROBLEMS WITH DEDICATION CONSTRAINTS: COMBINING SIMULATION AND CAPACITY PLANNING

Andreas Klemmt
Gerald Weigert

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

ABSTRACT

The main idea of the presented new approach is to join a discrete event simulation (DES) and mathematical programming techniques (i.e. mixed integer programming, MIP) for optimization of complex manufacturing processes. Thereby, a DES model allows a detailed problem description. For a target oriented optimization several capacity allocation problems are solved by a MIP solver, reducing the degrees of freedom in the DES model. As an example a typical parallel machine scheduling problem arising in semiconductor industry was chosen. Different process constraints like machine dedications, setups, auxiliary resources and processing time dependences are discussed – advantages and disadvantages of simulation-based and exact scheduling approaches are drafted. The investigated optimization goals comprise the reduction of total tardiness and setups efforts as well as a balanced machine utilization. Based on real manufacturing data of a wafer test area this approach is evaluated.

1 INTRODUCTION

Scheduling approaches in semiconductor manufacturing have been published for more than ten years. Most of them are special solutions. Up to this day there has been no general solution which answers the question "how to optimize." There is no commercial product available yet that is ready to use for the daily scheduling tasks without modifications. One of the reasons for this is that most of the practice-oriented scheduling tasks are NP-hard optimization problems (Brucker 2004). Hence, for solving complex scheduling problems, a lot of heuristics and decomposition methods were developed and investigated. A comprehensive overview about several of such approaches can be found in Ovacik and Uzsoy (1997) or Gupta and Sivakumar (2002). Thereby, problem-specific heuristics in combination with simulation and scheduling systems have shown the best efficiency.

To model dynamic manufacturing systems with complex resource constraints and large problem sizes (hundreds of jobs and machines), DES systems are primarily used. As one of the first scientists, Sivakumar (1999) issued an online capable simulation model for test equipment groups based on automated model generation. Such online parameterized models are also described in Potoradi et al. (2002) or Horn et al. (2006). But especially when using simulation, not only as a parameter forecast instrument but also for online scheduling decisions, several optimization aspects are requested.

Basically, a DES system is not an optimization system. It operates time directed and based on priority rules. So, the benefit of using simulation is primarily the knowledge of "what would be happen …" if the rules in the DES model reflecting the rules in the modeled manufacturing system. To enhance a DES system to an optimization system, primarily the method of simulation-based optimization is used (Fu, Glover and April 2005). This technique allows a comparison of different schedules by simulating a model several times. Therefore, the DES model contains a set of control variables influencing the behavior of the DES

model. These control variables are modified by heuristic search algorithms (e.g. threshold or genetic algorithms). The crucial point of this approach is the definition of problem specific control variables and the number of iterations needed to improve the performance of the schedule. Often the search space, spanned by the control variables, is exponentially increasing; so, there is a large number of iterations needed, which contradicts the online scheduling ability. In this paper a new approach for simulation-based optimization is presented. It is based on the idea that instead of changing control variables, the structure of the model itself is optimized. Therefore, several capacity allocation problems are solved. As consequence, the model is simulated only a few times and the online scheduling ability is maintained.

The paper is organized as follows. In section 2 typical parallel machine problems are drafted which can be found – with partly different side constraints – in nearly every semiconductor fab. Simple examples motivate the problems of simulation-based (rule-based) schedule generation. In section 3 a first static capacity allocation problem is discussed, which allows an optimized decision making. In section 4 a practical problem arising in a wafer test area is presented. Also a combined approach of simulation and capacity optimization is explained in detail. First results of this approach are shown in section 5.

2 MOTIVATION AND PROBLEM DISCUSSION

From the viewpoint of a scheduling problem, semiconductor manufacturing can be considered as a flexible job shop with different side constraints. That means, every job has a route with a predefined number of operations. At every operation the job is processed at one machine out of a group of parallel machines – the workcenter. Typical workcenters in semiconductor manufacturing are lithography, etching, furnace, polishing, implantation and wafer test. Depending on the technological process performed at the workcenter, different side constraints for scheduling have to be regarded (setups, auxiliary resources, batch processing etc.). One constraint, which can be found at nearly every workcenter, is the so called dedication constraint. As a result, the jobs cannot be processed on every machine of the workcenter. The reason for this can be drafted as follows: typically every job is assigned at every operation with a so called recipe specifying technological process parameters. Often only a subset of the machines of a workcenter is qualified for a recipe. This results from qualification and installation efforts (e.g. resists, gases, correlations). In the $\alpha \mid \beta \mid \gamma$ notation of Graham et al. (1979) dedication constraints are marked as M_i in the β-field.

While traditional dispatching rules (Pinedo 2008) like SPT – shortest processing time, ATC – apparent tardiness cost or SST – shortest setup time perform very well at identical parallel machine groups (e.g. SPT is optimal for $P_m \mid\mid \Sigma C_i$), their performance is much poorer if dynamic aspects (e.g. release dates r_i) and dedications have to be regarded, too. Because dispatching rules are a fundamental part of a DES system, some typical problems of simulation-based schedule generation can be observed repeatedly. Two of them are shown below. Thereby, the descriptions made are solely based on workcenters.

Example 1: Parallel machine problems with dedication constraints

A problem instance of the type $P2 \mid M_i;\ r_i \mid \Sigma C_i$ is investigated. So, the observed workcenter consists of two parallel machines. The jobs have release dates r_i and processing times p_i (Figure 1).

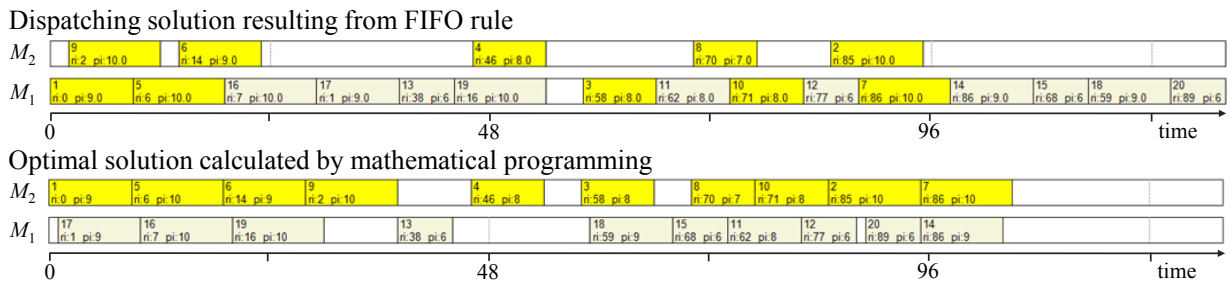


Figure 1: Parallel machine problem with dedication constraints

One half of the jobs (J_1 - J_{10}) belongs to one recipe that can be processed on both machines. The other jobs (J_{11} - J_{20}) belong to a second recipe (see coloring in Figure 1). They can only be processed on machine M_1. Figure 1 (upper part) shows the solution resulting from simulation. Thereby, all jobs are supplied concerning their release dates. If a machine becomes empty, the next available job is started – so this is a simple non-delay FIFO-rule. The lower part of Figure 1 shows the optimal solution of the described problem by using mathematical programming.

Example 2: Parallel machine problems with setup constraints

A problem instance of the type $P2 \mid r_i; p_i = 10; s_{ij} \mid \Sigma C_i$ is investigated. So, the observed workcenter again consists of two parallel machines. The jobs have release dates r_i and fixed processing times p_i . Furthermore, a setup is necessary between two jobs J_i and J_j if they not belong to the same recipe (coloring of the jobs). Initially one machine has the setup for recipe one and the other machine for recipe two. Figure 2 (upper part) shows the solution that results from the simulation. Thereby, the SST-rule is used. That means, if a machine becomes empty, the first job, which is matching the setup state of the machine, is started. So, if only jobs of the other recipe are available, a setup is performed. The lower part of Figure 2 shows the optimal solution of the described problem by using mathematical programming.

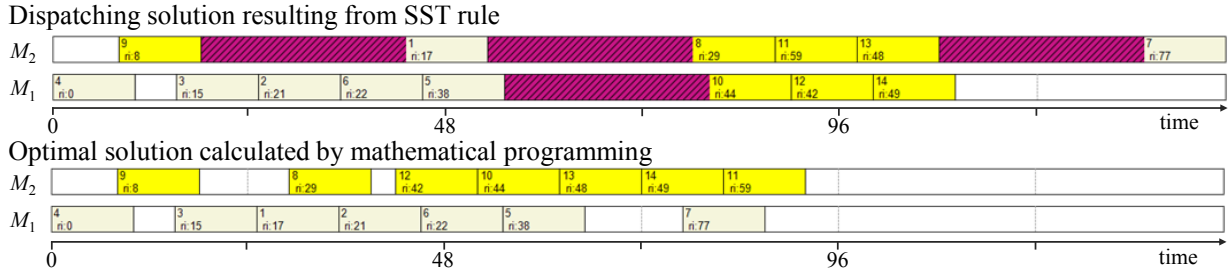


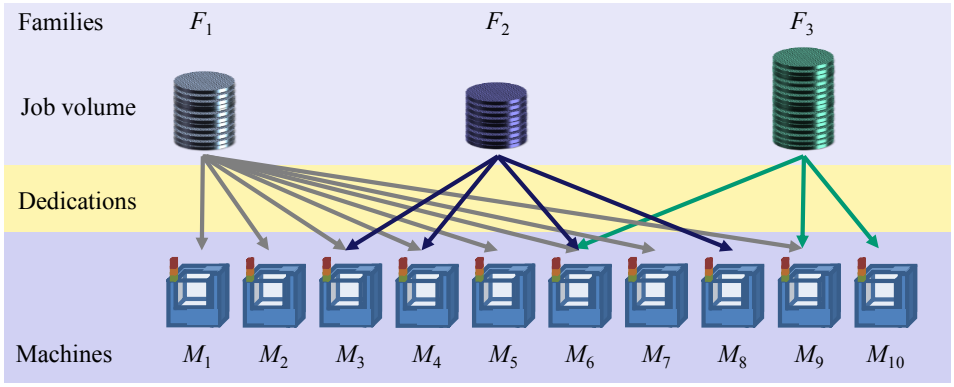
Figure 2: Parallel machine problem with setup constraints

The used dispatching rules in both examples are very simple. There are possibilities to improve the solution quality of the DES model (e.g. by limiting the number of identical setups). However, if the dedication scenario is more complex and release dates are unknown – because of previous operations – the problem lying behind can also be observed in practical application: the decisions made at discrete time events are uncoupled from the WIP (work in process) volume standing behind the recipes. A DES system does not tend to keep a machine idle, while there are still waiting jobs for it – even this results in efforts or unbalanced loadings. However, a dispatching solution like shown in Figure 2 would never be excepted by practitioners. For both examples, shown above, it is easy to see that a beneficial strategy is to process one recipe on one machine exclusively; that means to reduce the dedication scenario. The approach presented in the next sections aims on an automated calculation of potential reductions in the dedication scenario, to resolve capacitive unnecessary degrees of freedom. This will improve the simulation results.

3 CAPACITY ALLOCATION

For specific scheduling problems, capacity allocation methods can be used to limit unnecessary degrees of freedom in the dedication scenario or to resolve them. Dedication constraints occur in many practical problems and can lead – as shown in section 2 – to unfavorable schedules. This is primarily the case, if at defined events (control decisions in a DES model) a lack of knowledge about future job arrivals (and the WIP volume) exists or this knowledge is not regarded in the control decisions. Even if it is possible to improve the DES model, for example by calculating capacitive bounds for the maximal number of parallel setups, the question on which machines these setups are useful to perform is still not answered. The following model will help to answer this question.

Let m be the number of parallel machines M_k ($k = 1,\ldots,m$) at a workcenter and f the number of different families (recipes). Thereby, every job belongs to exactly one family F_i ($i = 1,\ldots,f$). Then a matrix $D \in \{0,1\}^{m \times f}$ specifies the dedication scenario (Figure 3). Furthermore, let $D_k := \{ i \mid D_{ki} = 1 \}$ the set of recipes allowed for processing on M_k, n_i the job volume of family F_i and $p_{ki} > 0$ ($D_{ki} = 1$) the processing time for a job of recipe F_i on Machine M_k.



$$D = \begin{array}{c} \\ M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \\ M_8 \\ M_9 \\ M_{10} \end{array} \begin{array}{c} \begin{array}{ccc} F_1 & F_2 & F_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

Figure 3: Dedication scenario at parallel machines

The goal is to calculate a reduced dedication matrix $D^{\text{red}} \in \{0,1\}^{m \times f}$, which allows a uniform load distribution on the tools and which locks families on tools, getting into potential overload situation. Thereby, it is (additionally) possible to define some boundaries $m_i^{\min}$ and $m_i^{\max}$, which specify the minimal and maximal number of machines to be qualified for recipe F_i, and a parameter $n_i^{\min}$, enforcing a minimal WIP, which is assigned to machine M_k, if $D_{ki}^{\text{red}} = 1$.

To calculate D^{red}, a simple capacity allocation problem has to be solved, which is widely discussed in literature for different types of assignment problems (Akcali, Üngör, and Uzsoy 2005). Therefore, the following decision variables of a mathematical model have to be defined:

$x_{ki} \in \mathbb{R}_+$	number of jobs from family F_i assigned to machine M_k; ($k = 1,\ldots,m$; $i \in D_k$),
$y_{ki} \in \{0,1\}$	family F_i is used on machine M_k, 0 otherwise; ($k = 1,\ldots,m$; $i \in D_k$),
$C_{\max} \in \mathbb{R}_+$	maximum workload.

Optimization model 1

$$C_{\max} \rightarrow \min \qquad \text{subject to} \tag{1}$$

$$\sum_{i \in D_k} x_{ki} = n_i \qquad k = 1,\ldots,m, \tag{2}$$

$$n_i y_{ki} \geq x_{ki} \qquad i = 1,\ldots,n;\ k \in A_i, \tag{3}$$

$$n_i^{\min} y_{ki} \leq x_{ki} \qquad i = 1,\ldots,n;\ k \in A_i, \tag{4}$$

$$\sum_{i \in D_k} y_{ki} \geq m_i^{\min} \qquad k = 1,\ldots,m, \tag{5}$$

$$\sum_{i \in D_k} y_{ki} \leq m_i^{\max} \qquad k = 1,\ldots,m, \tag{6}$$

$$\sum_{i \in D_k} p_{ki} x_{ki} \leq C_{\max} \qquad k = 1,\ldots,m. \tag{7}$$

Objective function (1) optimizes the load balancing on the machines. This is reached by minimizing the maximum workload on the machines (7). The workload is the amount of working time for processing

all jobs assigned to a machine (via x_{ki}). Equation (2) forces that all jobs are planned. Constraint (3) ensures that, if $x_{ki} > 0$ then follows $y_{ki} = 1$. Vice versa, equation (4) indicates that $x_{ki} > n_i^{\min}$ if $y_{ki} = 1$. Equations (5) and (6) ensure the minimal and maximal number of machines, which are qualified for each recipe.

Optimization model 1 is a pure capacity allocation problem. So, it cannot solve a scheduling problem because all dynamic aspects are ignored. Only for the special case of the $P_m \mid M_i \mid C_{\max}$ problem, an optimal schedule can be derived from the optimal solution of model 1 by ignoring y_{ki}, $m_i^{\min}$, $m_i^{\max}$ and $n_i^{\min}$ and enforcing x_{ki} to be integer. That means, $C_{\max}$ is then the minimal makespan for this problem.

For non-static problems, $C_{\max}$ can be regarded as lower bound for makespan, which implies an optimized load balancing. The result of the model is a reduced dedication matrix D^{red} with $D_{ki}^{\text{red}} = 1$, if $y_{ki} = 1$, $D_{ki}^{\text{red}} = 0$ otherwise ($k = 1,\ldots,m$; $i \in D_k$). Then, for the example shown in Figure 3, D^{red} can be equal to:

$$D_1^{\text{red}} = \begin{array}{c} \\ M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \\ M_8 \\ M_9 \\ M_{10} \end{array} \begin{array}{c} \begin{array}{ccc} F_1 & F_2 & F_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ \mathbf{0} & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{array}, \quad D_2^{\text{red}} = \begin{array}{c} \\ M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \\ M_8 \\ M_9 \\ M_{10} \end{array} \begin{array}{c} \begin{array}{ccc} F_1 & F_2 & F_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \mathbf{0} & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \text{ or } D_3^{\text{red}} = \begin{array}{c} \\ M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \\ M_8 \\ M_9 \\ M_{10} \end{array} \begin{array}{c} \begin{array}{ccc} F_1 & F_2 & F_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \mathbf{0} & 1 & 0 \\ \mathbf{0} & 1 & 0 \\ 1 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \mathbf{0} & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

depending on the choice of $m_i^{\min}$, $m_i^{\max}$ and $n_i^{\min}$. So, it is possible to control the density of D^{red} with the help of these three parameters. For problems with expensive setups (Figure 2), it is appropriate to get a reduced dedication matrix with a low number of different recipes on each machine, as shown by D_3^{red}. A more dense matrix D^{red} (as shown by D_1^{red} or D_2^{red}) leads to more degrees of freedom and to flexibility – especially concerning due dates and completion dates. Now, for calculating a schedule out of D^{red}, the DES system is used. Thereby, the main idea of the presented approach in this paper lies in an iterative improvement of D^{red} by combining simulation and the solving of capacity allocation problems.

Model 1 can be extended by several other constraints as setup conditions, batch processing, auxiliary resources and availability constraints, too. Some of them are discussed in the next section on the example of semiconductor manufacturing problem, arising in wafer test area.

4 APPLICATION EXAMPLE WAFER TEST

The type of the scheduling problem, which is arising in wafer test area, is $FJ \mid M_i; r_i; d_i; s_{ij}; aux; rcrc \mid \Sigma T_i$. That means a flexible job shop with parallel machines, release dates, recirculation, setups, dedications and auxiliary resources. The investigated performance measures are total tardiness and total setup time. The main problem in the wafer test is the efficient allocation of limited auxiliary resources (probe cards) to the machines (testers) with regard to the requirements of the expected WIP. A job can be processed on different machines with partly different probe cards. In addition, significantly varying processing times exist for different tester-to-probe card assignments. So, a lot of alternative processing possibilities have to be taken into account. Furthermore, machine setups have to be regarded if probe cards are changed.

For scheduling jobs with sequence dependent setup times on parallel machines, Lee and Pinedo (1997) developed a dispatching rule, which is called apparent tardiness cost with setup (ATCS). Pfund et

al. (2008) extended this ATCS-rule to consider release dates of the jobs. Scheduling approaches, primarily focusing on wafer test, can be found in Pearn et al. (2002) or Bang and Kim (2011). Thereby, Pearn et al. (2002) transforms parallel machine scheduling problems, which are arising in wafer test, into vehicle routing problems with time windows – the investigated performance measure is the minimization of total setup time. An approach for minimizing the total tardiness in a wafer test facility is presented by Bang and Kim (2011). They used a bottleneck detection and time window heuristics. But all wafer test approaches presented before, more or less neglect the allocation of auxiliary resources, which is a very crucial side constraint in the investigated facility. In the literature, auxiliary resource problems are, for instance, discussed in connection with reticle allocation problems in the photolithography area (Cakici and Mason 2007). In Klemmt et al. (2011) the concept of coupling of a DES system with a solver for capacity allocation problems was discussed from a practical viewpoint. This paper extends this approach.

4.1 General problem description

The main challenge of wafer test scheduling lies in the complexity of the dedication matrix D as well as in the setup and processing time constraints. The following points give a briefly problem description:

1. *Tester-to-probe card combination:* A test is to be performed by a single probe card, taken out of a set of feasible probe card types. For every probe card type there is a defined set of testers, which are capable to handle it. Furthermore, probe cards of the same type are only available in a limited number.
2. *Setups:* Tests are performed with different levels of temperature. This leads to job recirculation on the testers. For changing probe cards or temperature levels, a setup is necessary. Depending on the setup type, the length of the arising setup time differs.
3. *Probe card speeds:* On every touchdown, a probe card may test a number of chips. Therefore, the DUT (devices under test) stands for the number of simultaneously tested chips. Different probe card types can have different DUT, which approximately correlates with their testing speed.

Every job has at least one functional test operation in its route. If two operations of different jobs are identical (identical dedication scenario, equal processing times etc.), then they belong to one group. Because the test can be performed by different probe card types (different recipes), this group is called operation assignment O_i ($i = 1,\ldots,f$) in the following (Figure 4).

Afterwards, it is described how the problem is modelled with the help of a DES system. Then, it is demonstrated how optimization model 1 can be adapted to the new side constraints. Finally the coupling of both modelling approaches is discussed.

4.2 The DES Model

A DES model, containing all information concerning the current jobs stock, the incoming jobs (WIP) of the next week, the machines, the probe cards and the routes, is automatically created from the underlying MES-system. As a simulation engine, the simcron MODELLER is used. Figure 5 shows the implementation of the dedication scenario on the bases of an example route, which is containing operation assignment O_1 and O_2, as shown in Figure 4. The route contains two test operations. In operation one the job can be tested with a probe card of type T_1 on machine M_1 - M_7 or with a probe card of type T_2 on Tester M_8, M_9 or M_{10}. This is modeled by cascading OR- and AND-branches (Figure 5) in the DES system. Thereby, a test with probe card type T_1 is four times faster than the test with T_2. In a second test operation the job has to use T_2 solely. This test can then only be performed on M_9 or M_{10} (operation assignment O_2).

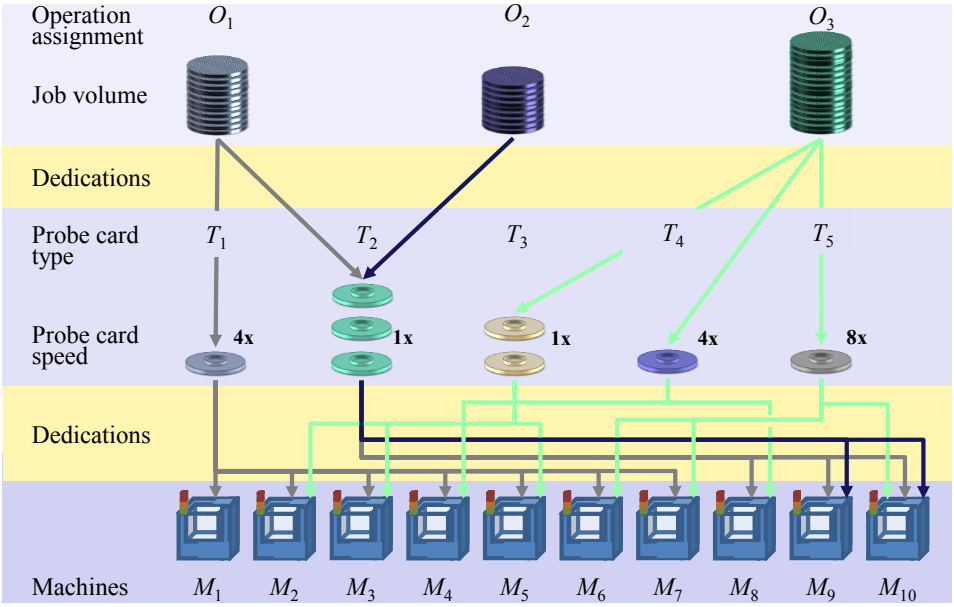


Figure 4: Dedication scenario and DES example route

OR – Branch:
One of the alternatives has to be choosen

AND – Branch:
Both, probe card and tester must be available

OR – Branch:
One of the testers has to be choosen

Delay step:
Process steps not belonging to the bottleneck workcenter are skipped in the model

Pocessing time:
Time for processing the job at the operation (depending on the probe card used)

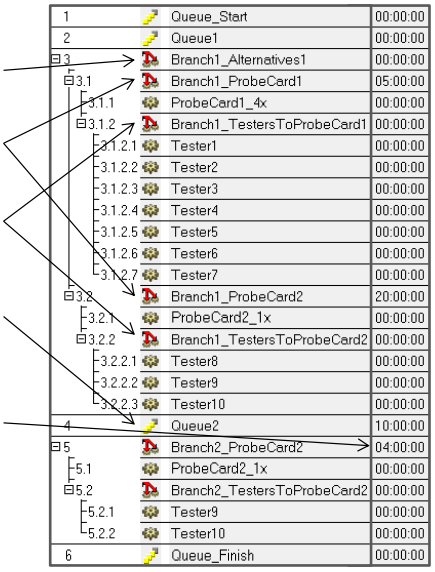


Figure 5: Dedication scenario and an example route of the DES system simcron MODELLER

All jobs have a release date and a due date. Operations, which are not belonging to the functional test, are mapped as delay steps in the DES model (see Queue2 in Figure 5). Furthermore, the DES model accurately reproduces the setup requirements of the manufacturing system. In contrast to Figure 5, the investigated practical problem contains several hundred jobs and dozens of machines, probe cards and different operation assignments. Consequently, there is a high diversity of variants for testing the jobs. For sequencing the jobs, a priority rule related to the ATCS-rule from Lee and Pinedo (1997) is used. That means, slack times, setup times and job priorities are considered at a dispatching decision. The simulation of the model takes several seconds (less than 30s for a simulation horizon of one week, dependent on the WIP). The result of the simulation run is a detailed production schedule. However, within this schedule the problems drafted in section two can be observed. Enhancing the approach by simulation-based opti-

mization is not beneficial in this application scenario. The reason for this lies in the complex dedication constraints, which makes it hard to find and implement some influential control variables. An enormous number of different tester-to-probe card assignments exists, which makes the search space of a simulation-based optimization not even approximately explorable.

4.3 The capacity allocation problem

The capacity allocation model 1 for reducing the dedication matrix (section 3) should be adapted to the new application scenario. Again, let m being the number of machines M_k ($k = 1,\ldots,m$), p the number of probe card types T_j (j=1,…, p) and f different operation assignments O_i ($i = 1,\ldots,f$). To enhance the model 1 to handle these conditions, let n being the number of different tester-to-probe card assignments A_l (l =1,…,n). Then, the dedication scenario is described by a binary matrix $D \in \{0,1\}^{n\times f}$. For every possible tester-to-probe card assignment also a processing time p_{li} (per job) exists, which is a element of a matrix $P \in \mathbb{R}^{n\times f}$. Next to the dedication scenario several other constraints have to be regarded, too. A parameter v_k specifies the (capacitive) availability of machine M_k ($k = 1,\ldots,m$). Another parameter c_j specifies the (capacitive) availability of probe card type T_j, that means the number of probe cards from type T_j multiplied by the length of the planning horizon. Also, a matrix $S \in \{0,1\}^{m\times p}$ is specified, if a probe card of type T_j is currently installed on machine M_k. Furthermore, a number n_i of jobs exists, which have operation assignment O_i in their route ($i = 1,\ldots, f$). At least two surjective functions f and g exists, which map each tester-to-probe card combination A_l ($l = 1,\ldots,n$) to exact one tester M_k ($k = 1,\ldots, m$) and one probe card type T_j (j = 1,…, p). That means $f(T_1 - M_2) = M_2$, $g(T_5 - M_{10}) = T_5$ etc. for the drafted problem. In addition, the following decision variables for the adapted optimization model 2 have to be defined:

$x_{li} \in \mathrm{N}_0$ … number of jobs for combination A_l and operation assignment O_i (l=1,…,n; $i \in D_l$),
$y_l \in \{0,1\}$ … tester-to-probe card combination A_l is used, 0 otherwise (l =1,…,n),
$C_{\max} \in \mathbb{R}_+$ … maximum workload.

Optimization model 2

$$C_{\max} \to \min \qquad \text{subject to} \tag{8}$$

$$\sum_{i\in D_l} x_{li} = n_i \qquad l = 1,\ldots,n, \tag{9}$$

$$y_l \sum_{i=1}^{f} n_i \geq \sum_{i\in D_l} x_{li} \qquad l = 1,\ldots,n, \tag{10}$$

$$y_l \leq \sum_{i\in D_l} x_{li} \qquad l = 1,\ldots,n, \tag{11}$$

$$\sum_{l=1(f(l)=k)}^{n} \sum_{i\in D_l} x_{li} p_{li} \leq v_k \qquad k = 1,\ldots,m, \tag{12}$$

$$\sum_{l=1(g(l)=j)}^{n} \sum_{i\in D_l} x_{li} p_{li} \leq c_j \qquad j = 1,\ldots,p, \tag{13}$$

$$y_l = 1 \qquad j = 1,\ldots,p; k = 1,\ldots,m; S_{kj} = 1; l = 1,\ldots,n; f(l) = k;\ g(l) = j, \tag{14}$$

$$\sum_{l=1(f(l)=k)}^{n} \sum_{i\in D_l} x_{li} p_{li} \leq C_{\max} \qquad k = 1,\ldots,m. \tag{15}$$

The model can be interpreted as follows. Objective function (8) minimizes again the maximum workload on the machines, restricted by equation (15). Constraints (9), (10) and (11) are an adaption of constraint (2), (3) and (4). Thereby, the additional boundaries for controlling the density of D^{red} are initially

ignored – only x_{li} is forced to be integer. With the help of equation (12), the machine availability boundaries are regarded. Constraint (13) restricts the probe card availabilities. Equation (14) enforces the initial probe card setup. This model is extended by several additional constraints in the following.

4.4 Coupling simulation and capacity allocation

The DES model allows a detailed description of the manufacturing problem (hundreds of jobs and dozens of machines) with all of its process constraints. However, as a consequence, this detailed model has an enormous diversity of variants not even approximately to cope with traditional approaches (e.g. simulation-based optimization). The basic idea of the new approach is not, to define control variables for changing parameters in the model, but rather to optimize the model itself. All priority rules or other control instructions in the DES model keep untouched. Only the degrees of freedom for tester-to-probe assignments are reduced by solving capacity allocation problems. The goal is, to find efficient assignment of probe cards to testers (so called dedication corridors) for predefined time horizons. This should help to withdraw the opportunity of a dispatching rule, to choose a tester-to-probe assignment that has contra-productive effects on the overall system. For solving the capacity allocation problem and for reducing the degrees of freedom in the DES model, it is coupled with a MIP solver. Thereby, the DES model is optimized in seven steps which are drafted in the following:

Step 1 – Simulation: The DES model is built automatically from the underlying MES-system with all restrictions. The model is simulated once (with priority rule ATCS). All described parameters for the optimization model 2 are extracted. The result of step one is an initial schedule with corresponding manufacturing parameters (total setup time, total tardiness, tester utilization etc.).

Step 2 – Optimization: Optimization model 2 is solved with objective function (8). The result is a dedication matrix D^{red} with $D_{li}^{\text{red}} = 1$, if $x_{li} > 0$, $D_{li}^{\text{red}} = 0$ otherwise ($l = 1,\ldots,n$; $i \in D_l$). Thereby, a capacitive load balancing bound $C_{\max}$ is calculated. To reach this bound, a maximized utilization of probe cards with a high number of DUT is required.

Step 3 – Simulation: The bound $C_{\max}$ is – as mentioned in section 3 – only a theoretical lower bound for the makespan, which implies a good balancing. This is due to the fact that the capacity allocation model neglects any dynamic aspects of the manufacturing system (e.g. job release dates, batches, which are resulting from the same setup strategy). Hence, the results of optimization model 2 are used as an input for the DES model. Thereby, the reduced dedication matrix D^{red} is used to disable unnecessary degrees of freedom in the DES model.

Figure 6 shows a representing example related to the problem, discussed in the Figure 4 and 5. Afterward, a simulation of the updated DES model is performed. The result of this simulation run is the makespan $C_{\max}^{*}$, which is also reachable in the dynamic system.

Step 4 – Optimization: A modified version of the optimization model 2 is solved. Therefore, a new decision variable has to be defined:

$z_k \in \{0,1\}$ … at least one probe card is assigned to machine M_k, 0 otherwise ($k = 1,\ldots,m$).

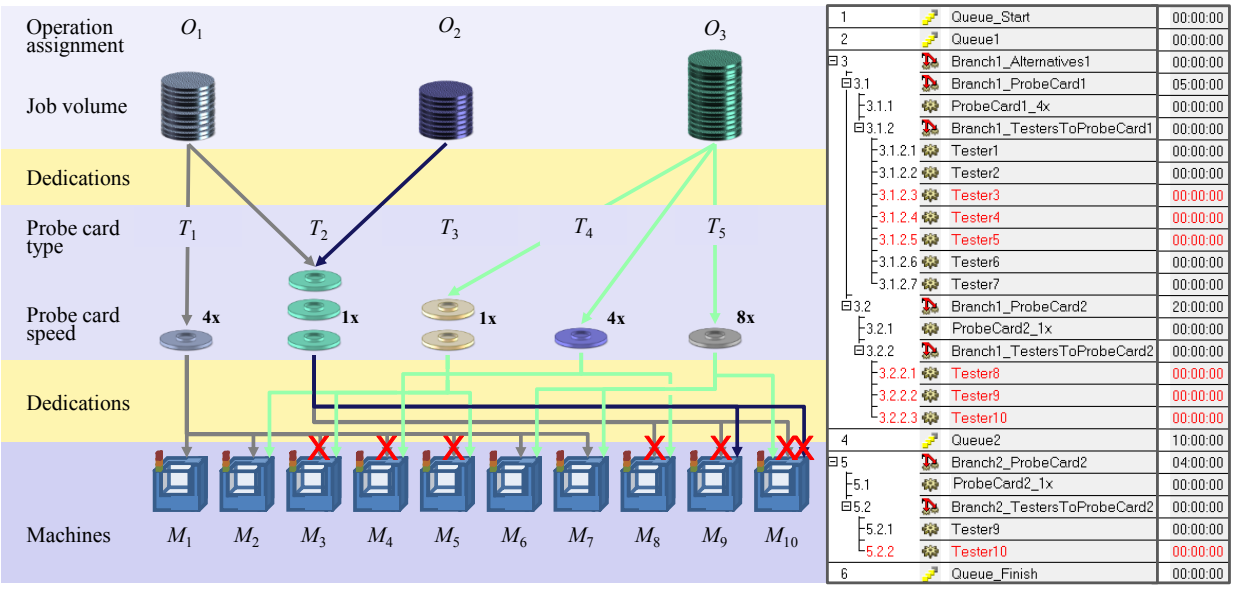


Figure 6: Reduced dedication scenario and resulting DES example route

Optimization model 3

$$\omega_1 \sum_{l=1}^{n} y_l - \omega_2 \sum_{k=1}^{m} z_k + \omega_3 \sum_{l=1}^{n} \sum_{i \in D_l} x_{li} p_{li} \rightarrow \min \qquad \text{subject to} \tag{16}$$

$$\sum_{l=1(f(l)=k)}^{n} \sum_{i \in D_l} x_{li} p_{li} \leq C^*_{\max} \qquad k = 1,\ldots,m, \tag{17}$$

$$\sum_{k=1}^{m} z_k \geq m^{\min} \tag{18}$$

$$z_k \leq \sum_{l=1(f(l)=k)}^{n} y_l \qquad k = 1,\ldots,m, \tag{19}$$

$$z_k m \geq \sum_{l=1(f(l)=k)}^{n} y_l \qquad k = 1,\ldots,m. \tag{20}$$

In this model, the values for the maximum machine availability boundaries v_k are reduced to $C^*_{\max}$ ($k = 1,\ldots,m$). So, by constraint (17) only solutions are further regarded allowing an optimized usage of probe cards with a high number of DUT. The goal of step 4 is the minimization of probe card setups. Thereby, the setups are not explicitly modeled. Rather, objective function (16) minimizes the total number of probe card to tool assignments and maximizes the probe card spread. Furthermore, the total workload is minimized. Thereby, the different goals are weightable by ω_1, ω_2 and ω_3. Equation (18) forces a minimum probe card spread. Restriction (19) and (20) bounding constraints for z_k.

Step 5 – Simulation: The result of step 4 is again a reduced dedication matrix D^{red}. Some alternatives in the DES model are now disabled – regarding to D^{red}. The result is that the number of setups, in a validating simulation run, is significantly decreased. The reason for this lies in the fact that the DES system does not need to choose between numerous different probe card types for one tester. On the other hand, necessary setups for high priority jobs or delayed jobs do not occur. So, after the simulation of the DES model has finished, all operation assignments O_i ($i = 1,\ldots,f$) with such violations are marked.

Step 6 – Optimization: In step 6 the MIP model of step 4 is further extended. Thereby, primarily one constraint is attached, enforcing all marked assignments O_i ($i = 1,\ldots,f$) to get a defined number of minimal assignments (alternatives). This is done by adding constraint (21) to model 3:

$$x_{li} \le n_i^{\max} \qquad l = 1,...,n;\ i \in D_l. \tag{21}$$

Thereby, $n_i^{\max}$ limits the maximal allocable load for each tester to probe card assignment, which implicitly enforces a load distribution on a defined minimal number of machines.

Step 7 – Simulation: The result of step 6 is again a reduced dedication matrix D^{red}. Now all jobs with critical due dates or high priorities have more degrees of freedom in the DES model as in the simulation of step 5. A final simulation run delivers an optimized schedule with corresponding manufacturing parameters (total setup time, total tardiness etc.).

5 RESULTS

The seven step optimization approach was tested in the course of an off-line study, where real data of a testing process in a wafer fab were used. The result is a detailed production schedule for a time horizon of approximately one week. Due to this schedule, the result of a DES model regards all existing process constraints of the wafer test facility without simplifications. The schedule also delivers an optimized tester-to-probe card assignment with regard to the requirements of the incoming WIP, which is a crucial information. The presented method significantly improves the performance parameters of the initial schedule (Figure 7). To illustrate the relative changes, the performance parameters of the optimized schedule (after step 7) are scaled to the values of the initial schedule. As one can see, probe cards with high DUT are more effectively used (approx. 10% improvement). This leads to a decreased workload on the machines. In addition, the number of setups could be reduced significantly, which further lowers the machine utilization. This primarily results from the efficient reductions in the dedications. However, even with these reductions, the reached values for total tardiness have not worsened. This mainly results from the alternatives in the dedication scenario, which is still available for delayed jobs.

The single models, inclusive the DES models, are completely derived from the real production data, so an online-application could be developed easily. This approach is not only applicable for the wafer test, but, because of its generic qualities, is also applicable in other parts of semiconductor manufacturing. Some further details are described in Klemmt et al. (2011).

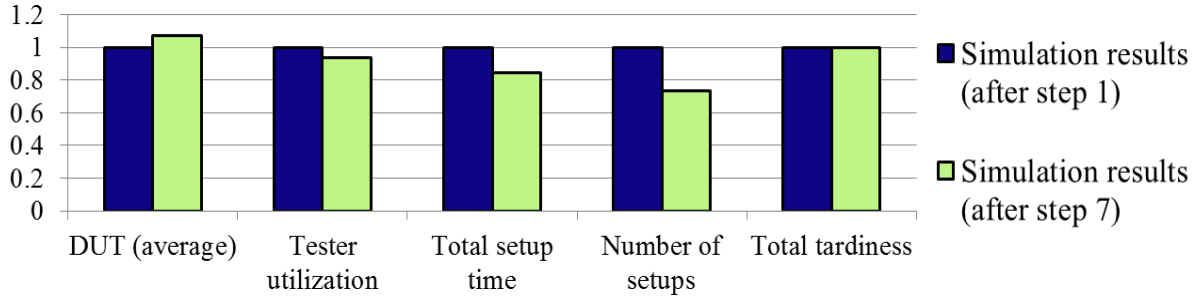


Figure 7: Performance results

6 SUMMARY

In the past simulation and mathematical methods were often competing offers for planning and control of manufacturing processes. But it has shown that both traditional methods – discrete event simulation as well as mathematical programming (i.e. mixed integer programming) – as a single application are not

suitable to solve more complex problems. Only the combination of both has the ability to solve a practical problem with an adequate number of resources and jobs. The principle is the alternate use of simulation and mathematical calculation to use the advantages of both methods and to reduce the effort drastically, especially the time, which is necessary for the optimization. So, the online application for the operational planning becomes possible, even for a lot of practical problems, where the calculation time should be not greater than a few minutes in general. First results from data of a wafer test facility confirm this statement. We are sure that the general idea is applicable for a lot of other scheduling problems, not only in semiconductor manufacturing.

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 and 13140/2171).

REFERENCES

Akcali, E., A. Üngör, and R. Uzsoy. 2005. "Short-term capacity allocation problem with tool and setup constraints." *Naval Research Logistics*, 52, 754-764.

Bang, J.-Y., and Y.-D. Kim. 2011. "Scheduling algorithms for a semiconductor probing facility." *Computers & Operations Research*, 38, 666-673.

Brucker, P. 2004. *Scheduling algorithms*. Springer.

Cakici, E., and S. J. Mason. 2007. "Parallel machine scheduling subject to auxiliary resource constraints." *Production Planning & Control*, 18, 217-225.

Fu, M. C., F. W. Glover, and J. April. 2005. "Simulation optimization: a review, new developments, and applications" In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 83-95. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Graham, R. L., E. L. Lawler, J. K. Lenstra, and A. H. G. Rinnooy Kan. 1979. "Optimization and approximation in deterministic sequencing and scheduling: a survey" In *Annals of Discrete Mathematics*, Vol.5, 287-326.

Gupta, A. K., and A. I. Sivakumar. 2002. "Simulation based multiobjective schedule optimization in semiconductor manufacturing" In *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yücesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, 1862-1870. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Horn, S., G. Weigert, S. Werner, and T. Jähnig. 2006. "Simulation based scheduling system in a semiconductor backend facility" In *Proceedings of the 2006 Winter Simulation Conference*, edited by L. R. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 1741-1748. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc..

Klemmt, A., J. Lange, G. Weigert, E. Beier, and S. Werner. 2011. "Combination of simulation and capacity optimization for detailed production scheduling in semiconductor manufacturing" In *Proceedings of the 21th International Conference on Flexible Automation and Intelligent Manufacturing*.

Lee, Y. H., and M. Pinedo. 1997. "Scheduling jobs on parallel machines with sequence dependent setup times" In *European Journal of Operational Research*, Vol.100, 464-474.

Ovacik, I.-M., and R. Uzsoy. 1995. "Rolling horizon procedures for dynamic parallel machine scheduling with sequence dependent setup times" In *Journal of Production Research*, 33, 3173-3192.

Ovacik, I. M., and R. Uzsoy. 1997. *Decomposition methods for complex factory scheduling problems*. Kluwer Academic Publishers.

Pearn, W. L., S. H. Chung, M. H. Yang, and A. Y. Chen. 2002. "Minimizing the total machine workload for the wafer probing scheduling problem" In *IIE Transactions*, Vol.34, 211-220.

Pfund, M., J. W. Fowler, A. Gadkari, and Y. Chen. 2008. "Scheduling jobs on parallel machines with setup times and ready times" In *Computers and Industrial Engineering*, Vol.54, 764-782.

Pinedo, M. 2008. *Scheduling: theory, algorithms and systems*. Springer.

Potoradi, J., O. S. Boon, S. J. Mason, J. W. Fowler, and M. Pfund. 2002. "Using simulation-based scheduling to maximize demand fulfillment in a semiconductor assembly facility" In *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yücesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, 1857-1861. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sivakumar, A. I. 1999. "Optimization of a cycle time and utilization in semiconductor test manufacturing using simulation based, on-line, near-real-time scheduling-system" In *Proceedings of the 1999 Winter Simulation Conference*, edited by P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans, 727-735. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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 made his PhD in 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.

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.