Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, eds

IMPROVING FLOW LINE SCHEDULING BY UPSTREAM MIXED INTEGER RESOURCE ALLOCATION IN A WAFER TEST FACILITY

Dirk Doleschal
Jan Lange
Gerald Weigert

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

Andreas Klemmt

Infineon Technologies
Königsbrücker Straße 180

01099 Dresden, GERMANY

ABSTRACT

The effort for scheduling real manufacturing systems is generally very high for mathematical as well as for simulation-based methods. Combining both methods is the key for solving complex scheduling problems. The paper introduces a special approach, where at first a static resource allocation problem is solved by mixed integer programming (MIP). Based on the resulting reduced dedication matrices, feasible schedules are then generated by a discrete event simulation (DES). Possible applications can be found in many parts of the semiconductor manufacturing process, for example in the wafer test. The investigated wafer test consists of two pronounced bottlenecks; each of it is formed as a workcenter with its own dedication matrix. After testing the method with practice oriented benchmarks, the benefits of the approach are shown on data derived directly from the semiconductor manufacturing process.

1 INTRODUCTION

It is well known that most of the practice-oriented scheduling tasks are NP-hard optimization problems (Brucker 2004). So, often the only way remains in a discrete event simulation (DES), usually extended by a heuristic optimization component, to schedule dynamic systems with complex resource constraints and large problem sizes. As one of the first scientists Sivakumar (1999) issued an online capable simulation model for test equipment groups. As a prerequisite for practical application, the simulation models are generated automatically. Comparable simulation-based applications are for example also described in Potoradi et al. (2002) and Horn et al. (2006).

But especially when using DES not only as a parameter forecast instrument, but also for online scheduling decisions, the time-consuming aspect (hundreds to thousands of repeated simulation runs may be required) becomes more and more important. This is the reason why more and more applications make use of mathematical methods. However, this requires a decomposition of problems into smaller units. The goal is, to find a good solution for these single units, even if the sub problems are still NP-hard. 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 for instance 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.

In this paper, investigations for the wafer test in a semiconductor industry were carried out. The wafer test is located between the frontend and backend of a semiconductor manufacturing. Here, the chips on wafers are tested on functionality before they are separated and go to further processes in the backend. Because the wafer test is the direct connection between frontend and backend, it is important to fit due

978-1-4673-4782-2/12/$31.00 ©2012 IEEE

dates as well as possible to ensure a constant product flow. So, a smaller unit of the wafer test – the functional test – is considered. Thereby, the underlying scheduling problem is a single operation problem with unrelated parallel machines, release dates, setups and dedications. For these specific scheduling problems capacity allocation methods can be used to remove unnecessary equipment allocations in the dedication scenario. Dedication constraints are typical for many process steps in semiconductor manufacturing (cf. Klemmt and Weigert 2011) but also exist in other manufacturing processes. So the here presented methods can be easily adapted for other manufacturing processes. After the functional test was considered in the smaller model, the whole wafer test is scheduled with a Virtual Time Based Flow Principle (VTBFP) method in combination with a modified capacity planning algorithm. Therefore the effect of both methods is examined.

The paper is organized as follows: the process flow in the wafer test is shortly described in section 2. Because the functional test was recognized as a severe bottleneck in the flow line, in section 3 a single operation problem, based on special benchmark models, is investigated. These results are applied to practical data of an example wafer test in section 4. The paper is closed by a short summary.

2 WAFER TEST

The investigated manufacturing is a high-mix, low-volume facility. So, it is important to ensure a high throughput and machine utilization. Each product has its own routes through the wafer test and also dedication matrices exist. Because of the high variability in the products, the different routes and dedication matrices, it is hard to generate optimal schedules.

The wafer test consists of several process steps, where not all process steps are used by each product. In Figure 1 different product routes are shown exemplary for only a few process steps.

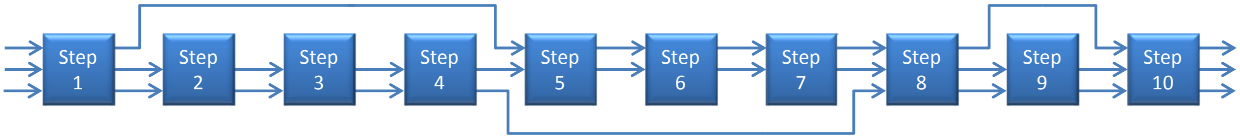


Figure 1: An example with only a few process steps and different routes

Thereby, some of the main steps can be:

- In/Out – These describes steps when the wafer arrive at the wafer test from the frontend or leave it to the backend.
- Functional test – This is a functional test, where the chips on the wafer were tested in parallel. This test need the highest amount of time in the wafer test. Therefore, a lot of parallel machines exist for this step, and not all machines are allowed for each product. The possible machine allocation is described in a dedication matrix. In contrast to the reality, where several functional tests exist, in this investigation only two functional tests were depicted. These process steps uses the same machines which also leads to reentrant scheduling problems.
- Furnace process – In this process step, aging of the chips is done to detect early failures.
- Cleaning process – The wafers are cleaned to be ready for further processing.
- Packaging – The wafers are prepared for transport to backend or customer.

The processing times vary widely. For example, the first and second functional test have partly very long and highly diffusing processing times for different products. Other process steps (for example in/out processes) may only have one tenth and less of the processing times arising in the functional test. These large differences result in a high potential for optimization. Also uncertainties in process time lead to challenges for scheduling. In case of the high difference in the process times for different products in the functional test, this operation is determined as a bottleneck in the wafer test, where here the bottleneck is defined as the process step with the highest average process time and highest utilization. Therefore, this workcenter should be considered separately and will be the point of interest in the next section.

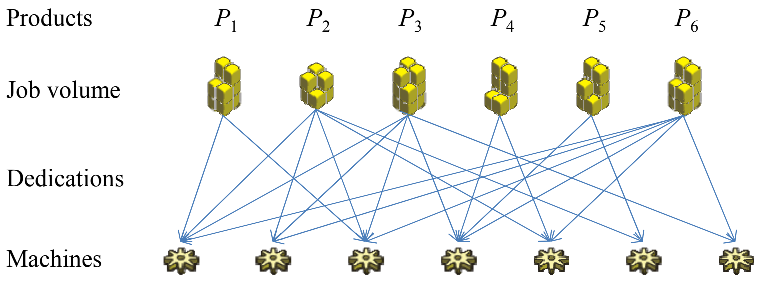
The objectives in the wafer test are a high throughput and hence a high machine utilization. Further objectives are the time based goals tardiness and lateness, which ensure to hold given due dates of the products. According to Pinedo (2008) the lateness of a job is defined as the difference between the due date and the completion time, which is positive if the job is late and negative if it is completed early. The tardiness only consists of the late component of the lateness, so early jobs have a tardiness of zero and late jobs a positive value. The reduction of necessary setups is also an important objective.

3 THE SINGLE OPERATION PROBLEM

In the real manufactory the functional test has proven to be a bottleneck in the wafer test. So, in this section the functional test is investigated more in detail and the effect of an upstream mixed integer based capacity planning is shown. The capacity planning is used to support an existing dispatching rule getting better results.

3.1 Problem Description

The basis for this problem is the functional test described in section 2. So, this is a single operation problem where the job is processed at one machine out of a group of parallel unrelated machines – the workcenter. Several side constraints for this workcenter exist. These are dedications, setup times, heterogeneous process times and release dates. In the three field α | β | γ notation by Graham et al. (1979) this problem can be written as $R_m \mid r_{ij}, s, p_{ij}, M_i \mid C_{\max}$. Thereby the α field describes the equipment area, the β field contains the process conditions and the γ field includes the optimization objective. In Figure 2 this is exemplary illustrated with six different product families and seven machines.



$$\begin{array}{c|ccccccc} & M_1 & M_2 & M_3 & M_4 & M_5 & M_6 & M_7 \\ \hline P_1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ P_2 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \\ P_3 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ P_4 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ P_5 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ P_6 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \end{array}$$

Figure 2: Schematic problem description including the corresponding dedication matrix

Thereby different product families with several jobs per family exist. Each family has dedications, which permit or disable a machine for this product. Each product family consists of a number of jobs which have to be processed on one or more of the permitted machines. The process times can be homogeneous or heterogeneous for all machines of a family. The machines can have setup times which occur when the family changes on the machine. Each job of a product has also a release date, so this job cannot be processed before this date.

According to the last section the objective is a high throughput, which is equal to a high utilization of the machines. This goal can be reached by minimizing the makespan (cycle time) $C_{\max}$. The makespan is equivalent to the maximum workload on the machines and so the optimization objective (which is considered) is a good load balancing over all machines. The number of setups is also an important objective in this case. So this is considered in the benchmarks and the results, too.

Now the challenge is to generate a valid schedule with the goal of a good load balancing on the machines. This means the utilization should be distributed equal over all machines. This problem is similar to an allocation problem, which was already described in (Doleschal, Lange, and Weigert 2012). So this method is used again, but with modified, more practical test instances. The changes of the benchmark are described in the next section. The creation of a schedule is done by a DES-system using the dispatch rule which is also described in the referred paper.

3.2 Test Instances

To test the described method under certain terms and conditions nearly 10,000 test instances were generated. Thereby the parameters, introduced in this section, are chosen in the way that they are close to the real process in the wafer test. The test instances are orientated on the problem description from previous section. The dedication matrix is chosen as followed:

The dedication matrix D is split in two types of jobs. One type has a high density, this means a high amount of allowed machines and the other type has a low density, which results in only a few allowed machines. So the dedication matrix D can be characterized with three parameters:

- The average percentage of products with low/high density, so this parameter describes the probability, whether a product dedication gets a low density respectively a high density
- The average number of allowed machines for products with a low density in the dedication matrix
- The average number of allowed machines for products with a high density in the dedication matrix

An example of a dedication matrix is shown in Figure 2. The test instances described in Table 1 were created and the effect of the mixed integer resource allocation method is calculated.

Table 1: Design of experiments (UD □ Uniformly distributed)

Parameter	**Values used**	**Total values**
Number of products n	10, 40, 100	3
Number of jobs per product n_i	UD ~ [50 100]	1
Number of machines m	5, 10, 20	3
Average amount of products with low density	0.2, 0.4	2
Low density	0.1, 0.3, 0.5	3
High density	0.9	1
Process time $p_{i,k}$	UD ~ [50 100]	1
Setup time s	0, 50, 200	3
Release date $r_{i,j}$	UD ~ [0 X*C_{min}]; $X \in \{0, 0.4, 1\}$	3
Homogeneous process times	true, false	2
	Number of independent instances	10
	Total problems	9,720

3.3 Results

With the help of the test instances, defined in the previous section, the effect of the mixed integer based resource allocation method is shown. For this, two schedules were generated for each test instance – one with the DES system and the original dedication matrix and one with the previous resource allocation, which generates a reduced dedication matrix. The used dispatching rule always generates non-delayed schedules, whereby products with a high amount of available jobs get a higher priority than products with only a few available jobs. The setup state is only changed if no further job of the current state is waiting.

The results are presented in Figure 3. Thereby, the average maximum workload ($C_{\max}$) from the DES system without previous capacity planning is normalized to value 1.0. The other results are pictured as a bar relative to this value. The first bar defines the result of the schedule, planned by the DES system after a capacity planning. The second bar is the lower bound for $C_{\max}$, calculated by MIP capacity planning. This is only a lower bound for the maximum workload $C_{\max}$, because no dynamic behavior is considered.

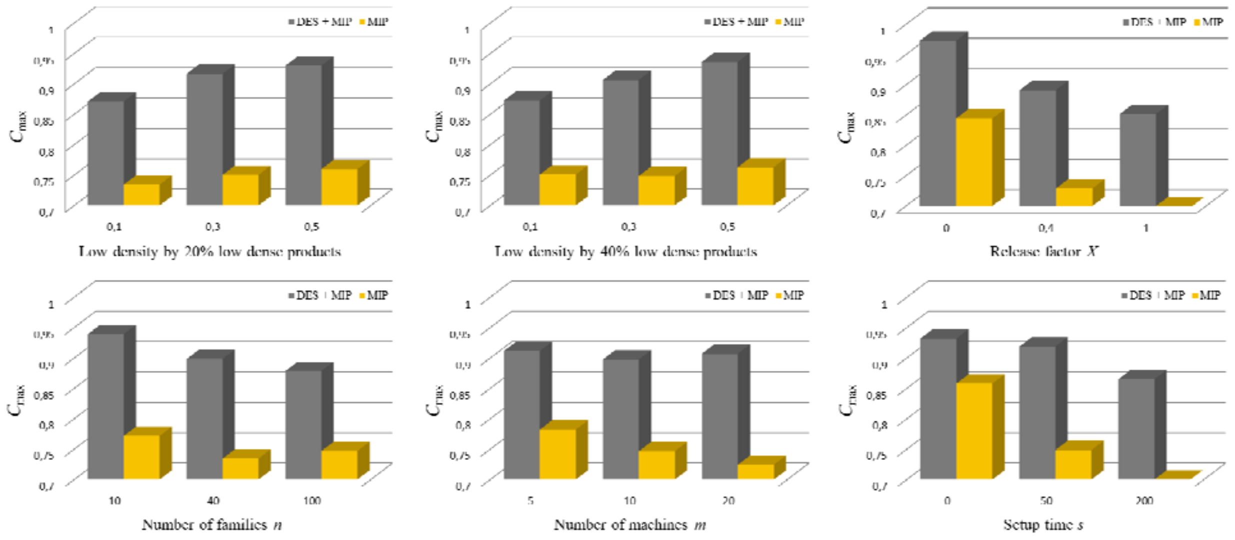


Figure 3: Results over all 9,720 benchmarks for objective $C_{\max}$

The upper left chart shows the results for all benchmarks with 20% low density products in relation to the parameter for the low density. Equivalent to this the upper middle chart shows the results for the benchmarks with 40% low density products. The chart in the upper right shows the results in relation to the release date. Thereby, the capacity planning has a bigger effect on non-static problems. The average improvement over all test instances is about 9.5%. Further, the diagrams in Figure 2 show the results in dependency of the number of families, number of machines or setup times.

The results show that the benefit from the presented mixed integer based capacity planning is relatively high. Thereby the time, needed by the capacity planning, is low (mostly less than 10 sec). So you get schedules which are in average 9.5 % better for the objective “maximum workload”.

The other considered objective is the number of setups. Because each setup generates cost and needs time, it is disadvantageous if the number of setups increases instead. In Figure 4 the results for this objective are shown in the same way as before. Naturally only these test instances with setup time $s > 0$ are considered.

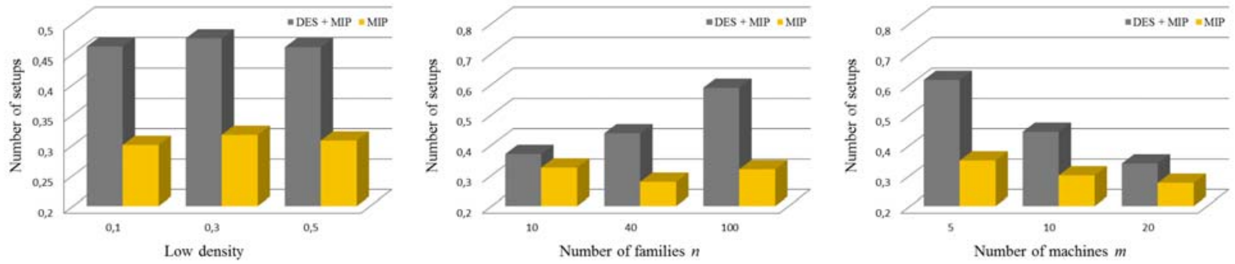


Figure 4: Results concerning number of setups over all 6,480 test instances with setup times

The value for MIP is simply the sum of all “1” in the dedication matrix D^{red}. The number of performed setups, generated by DES without MIP capacity planning, is normalized to 1.0 and the other results are shown relatively to this. In average, 53% less setups have to be performed after the upstream capacity planning by using the same dispatching rule. Naturally this cannot be transferred directly to the reality.

This MIP-based capacity planning shows a good result with low effort. So in the next section the whole wafer test is investigated and a MIP-based capacity planning for the VTBFP approach is introduced.

4 RESOURCE ALLOCATION WITH VTBFP

After the tests in section 3 have shown good results through the upstream mixed integer resource allocation method now the whole wafer test is investigated. For this, a modified resource allocation method is used to improve a clocked scheduling method – the Virtual Time Based Flow Principle (VTBFP). This method is not introduced in detail in this paper. Further information can be found in (Lange et al. 2012). Here only the used resource allocation method is described and the effect on scheduling is investigated.

The wafer test described in section 2 is implemented in a simulation model. For this the discrete event simulation system simcron MODELLER (Horn et al. 2006) is used. The data for this model is directly retrieved from a semiconductor manufactory. Thereby the simulation model includes 15 process steps and 14 different products. Two of the 15 process steps are operations of the functional test. In praxis, these products are divided into more time-critical products and filling ware. In addition 30 test machines are modeled, which are used in both functional tests.

In this paper, the focus lies in the functional test. These operations are the bottlenecks in this problem. So, a resource allocation is done for these two operations. Like described in section 2, these operations have dedication matrices for each product. Thereby, the dedication matrix can differ between the first and second functional test for a product. All processing times in the simulation model are stochastic. So uncertainties in processing times exist. These processing times were derived directly from the real test process. Because of the statistical certainty, 100 replications for each combination are done.

The objectives for the modeled wafer test orientate on those, defined in section 2. So, the investigated objectives are the throughput, the tardiness, the lateness and the number of setups. The throughput is the number of jobs completed within a week.

4.1 Function of Virtual Time Based Flow Principle

The basis for the capacity resource allocation method is the Virtual Time Based Flow Principle. This is a relatively new method (Keil et al. 2011). The main idea of this method is to create virtual barriers in a manufacturing system to meet due dates as good as possible. These barriers have the effect that they can delay jobs, inclusive priority jobs, if they reach the barrier too early. Thereby not all products must have these barriers. To distinguish these products, those products with barriers are called clocked products. In the dispatching rule of a scheduling system, these clocked products have a higher priority than the other products. In the current state, barriers are placed by an expert. Each job of a product gets its own time for a barrier, depending on the release date and the product of this job. In Figure 5 barriers for the three products of section 2 are placed as an example.

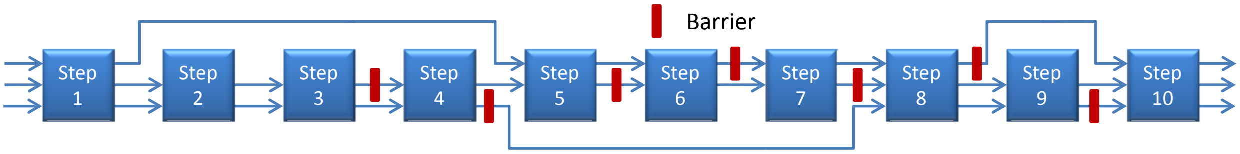


Figure 5: Barriers in an exemplary work flow with three products

In the theory of VTBFP clocked products should have machines exclusive in the best case or the dedication matrix should have as few overlap as possible with other products. More detailed information about this scheduling method can be found in (Lange et al. 2012).

4.2 MIP Model

Affected by the VTBFP method, a capacity planning model for the clocked products is created. The goal for this method is a reduced dedication matrix D^{red}, so that the clocked products get machines exclusively with as little overlap as possible in the dedication matrix. The following data is necessary to build the optimization model for the mixed integer programming model:

- n different product families F_i (i=1,…,n), where every product family includes n_i jobs,
- m different parallel machines M_k (k = 1,…,m),
- A dedication matrix $D \in \{0,1\}^{n \times m}$, which specifies permitted and disabled machines for each product. Also $D_k := \{\, i \mid D_{i,k} = 1\}$ is the set of products permitted for processing on machine M_k. In the same manner, $D_i := \{k \mid D_{i,k} = 1\}$ is the set of machines permitted for processing family F_i,
- $p_{i,k} > 0$ is the processing time for a job of product family F_i on machine M_k if $D_{i,k} = 1$,
- V_k is the availability for machine M_k, which is a capacity bound for this machine,
- a list L of priority (clocked) products, which includes all clocked products,
- parameter $n_i^{\min} \geq 0$ to describe the number of minimal jobs of product F_i that have to be processed on a used machine in the reduced dedication matrix D^{red}.

The parameter $n_i^{\min}$ is chosen by the user. Also the parameter V_k can be varied by the user to influence the denseness in the reduced dedication matrix. The other parameters where directly retrieved from the described scheduling problem. To calculate the reduced dedication matrix D^{red}, a capacity allocation problem has to be solved. Therefore the following decision variables of a mathematical model have to be defined:

$X_{i,k} \in \square^{+}$ number of jobs from family F_i assigned to machine M_k; (k = 1,…, m; $i \in D_k$),

$Y_{i,k} \in \{0,1\}$ product family F_i is used on machine M_k, 0 otherwise; (k = 1,…, m; $i \in D_k$),

$Z_k \in \square^{+}$ number of overlaps for machine M_k with clocked products; (k = 1,…, m).

The following mixed integer resource allocation model has the goal to minimize the overlap in the dedication matrix for the clocked products of the VTBFP. This means, the priority products (clocked products) should have machines exclusive, if possible. Using the defined data, the following optimization model can be formulated:

Optimization model 1

$$K \cdot \sum_{k=1}^{m} Z_k - \sum_{i=1}^{n} \sum_{k \in D_i} Y_{i,k} \to \min \qquad \text{subject to} \qquad (1)$$

$$\sum_{k \in D_i} X_{i,k} = n_i \qquad i \in \{1,...,n\} \qquad (2)$$

$$\sum_{i \in D_k} X_{i,k} \cdot p_{i,k} \leq V_k \qquad k \in \{1,...,m\} \qquad (3)$$

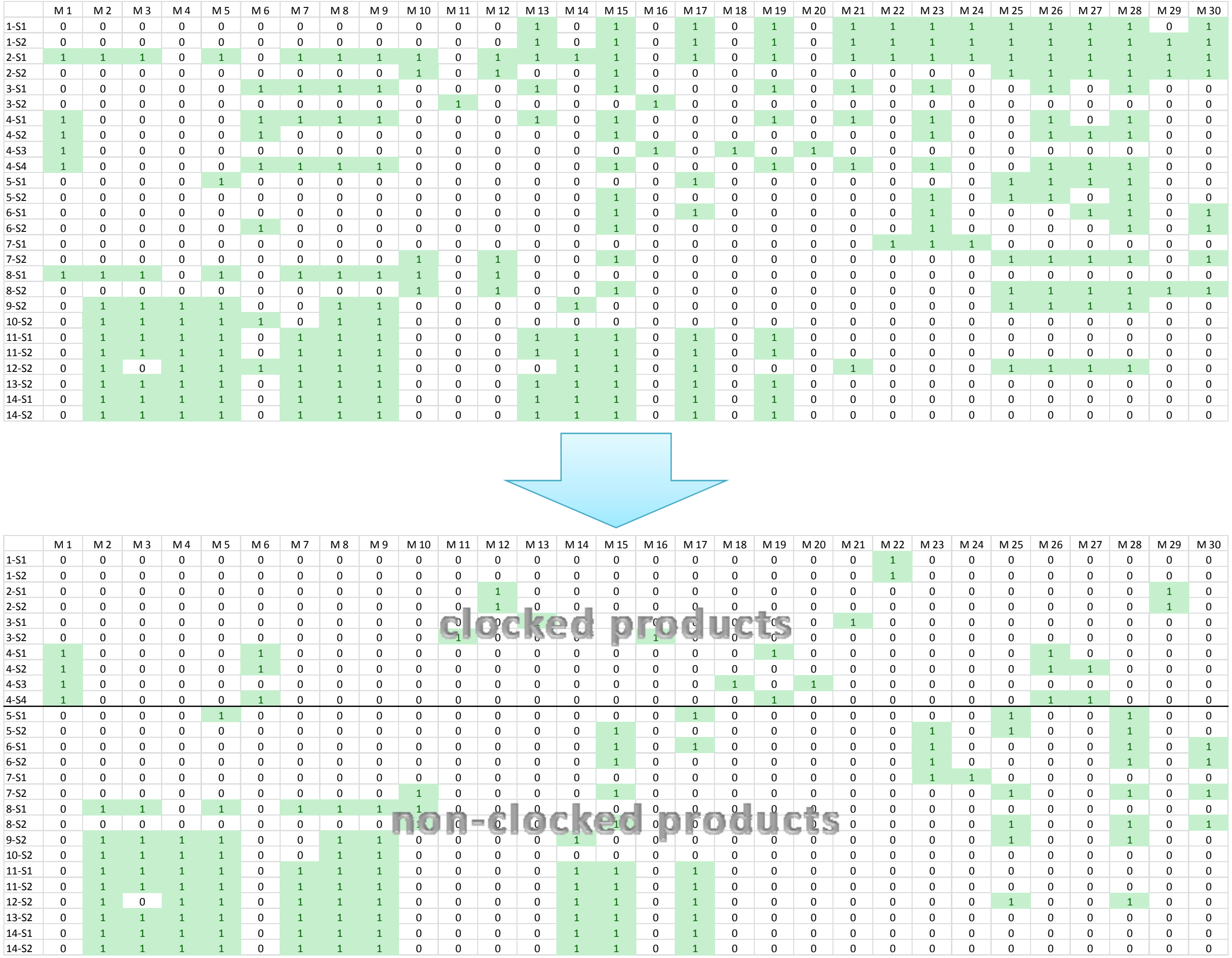
$$n_i \cdot Y_{i,k} \geq X_{i,k} \qquad i \in \{1,...,n\};\ k \in D_i \qquad (4)$$

$$n_i^{\min} \cdot Y_{i,k} \leq X_{i,k} \qquad n_i^{\min} > 0;\ i \in \{1,...,n\};\ k \in D_i \qquad (5)$$

$$K \cdot Y_{l,k} + \sum_{\substack{i=1 \\ i \neq l}}^{n} Y_{i,k} - Z_k \leq K \qquad l \in L;\ k \in D_l \qquad (6)$$

Subsequently, optimization model 1 has to be solved. Thereby, objective function (1) minimizes the machine overlap and maximizes the number of allowed machines in the reduced dedication matrix D^{red}. For this the constant K is a number, which is big enough to ensure that the reduction of overlap has the highest priority. In this case, K can be set as $K = m \cdot n$. Equation (2) forces that all jobs are planned. With the equation (3) the availability for the machines is used. So, this is a upper bound for the makespan. Constraint (4) forces that, if $X_{i,k} > 0$, than follows $Y_{i,k} = 1$. Vice versa equation (5) saves that $X_{i,k} \geq n_i^{min}$, if $Y_{i,k} = 1$. This is only used, if $n_i^{min} > 0$. The last equation (6) calculates the overlap for each priority product $l \in L$. This means, if product l is used on machine k ($Y_{l,k} = 1$), then the number of overlaps is the sum of all other products also planned on this machine. This equation also uses the big integer K, defined as $m \cdot n$.

In Figure 6 the function of this optimization model is shown on an example dedication matrix. Thereby the upper matrix shows the original dedication matrix, where each filled element represents a combination of an allowed product and machine. The products are on the vertical axis and the horizontal axis contains the test machines. In the lower dedication matrix, the upper jobs (above the line) are priority products, for which a machine should be planned exclusively (or nearly exclusively). Sometimes, there are two filled elements in neighbored lines of priority products on one machine (1-S1,1-S2 and M 22). This is, for example, the same product but with two different functional test steps, which are separated in the dedication matrix. So, this product has two product entries, one for the first functional test and one for the second functional test, but if both are planned on the same machine it is not counted as an overlap.



	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	M 11	M 12	M 13	M 14	M 15	M 16	M 17	M 18	M 19	M 20	M 21	M 22	M 23	M 24	M 25	M 26	M 27	M 28	M 29	M 30
1-S1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	0	1
1-S2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1
2-S1	1	1	1	0	1	0	1	1	1	1	0	1	1	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1
2-S2	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
3-S1	0	0	0	0	0	1	1	1	1	0	0	0	1	0	1	0	0	0	1	0	1	0	1	0	0	1	0	1	0	0
3-S2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4-S1	1	0	0	0	0	1	1	1	1	0	0	0	1	0	1	0	0	0	1	0	1	0	1	0	0	1	0	1	0	0
4-S2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0
4-S3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
4-S4	1	0	0	0	0	1	1	1	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	1	1	1	0	0
5-S1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	0
5-S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0
6-S1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1	1	0	1
6-S2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1
7-S1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
7-S2	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	1
8-S1	1	1	1	0	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8-S2	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
9-S2	0	1	1	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0
10-S2	0	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11-S1	0	1	1	1	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
11-S2	0	1	1	1	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
12-S2	0	1	0	1	1	1	1	1	1	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	1	1	1	1	0	0
13-S2	0	1	1	1	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
14-S1	0	1	1	1	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
14-S2	0	1	1	1	1	0	1	1	1	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0

	M 1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	M 11	M 12	M 13	M 14	M 15	M 16	M 17	M 18	M 19	M 20	M 21	M 22	M 23	M 24	M 25	M 26	M 27	M 28	M 29	M 30
1-S1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
1-S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2-S1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2-S2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3-S1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3-S2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4-S1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
4-S2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
4-S3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
4-S4	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
5-S1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0
5-S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0
6-S1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1
6-S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1
7-S1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
7-S2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1
8-S1	0	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8-S2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1
9-S2	0	1	1	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
10-S2	0	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11-S1	0	1	1	1	1	0	1	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
11-S2	0	1	1	1	1	0	1	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
12-S2	0	1	0	1	1	0	1	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0
13-S2	0	1	1	1	1	0	1	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
14-S1	0	1	1	1	1	0	1	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
14-S2	0	1	1	1	1	0	1	1	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 6: Example for MIP-based resource allocation (Top … original dedication matrix, bottom … partly clocked dedication matrix)

The density of the resulting dedication matrix can be influenced by the availability V_k of the machine M_k. This means, a low availability leads to a higher density in the reduced dedication matrix for the clocked products and so to a lower density in the area of non-clocked products. Vice versa a high availability has the effect that the clocked products get less machines, but the machines for non-clocked products were maximized, so the density in this area of the dedication matrix is higher.

4.3 Test setup

To investigate the effect of the described resource allocation method, several approaches with the VTBFP method and the mixed integer resource allocation method are performed. The VTBFP method is done for the described 15 process steps (including two functional tests) and 14 products. The products have different but deterministic process routes and stochastic process times, which vary for each product and step. More information can be found in (Lange et al. 2012).

In summary, 1,500 schedules were generated for one run. The products are clocked like described in Table 2, which means in scenario 1 no products are clocked and stepwise to scenario 15 more and more products are clocked, until in the last scenario all products are clocked. This was chosen because the priority in the manufacturing of the products increases in the same way. So the higher priority products should be clocked first. Non-clocked products either are normal products (product 1 – 9) or they are filling ware (product 10 – 14). In the simulation model the clocked products have the highest priority and the filling ware the lowest priority.

Table 2: Scenario table (N … normal product, F … filling ware, C … clocked product)

Scenario	Prod 1	Prod 2	Prod 3	Prod 4	Prod 5	Prod 6	Prod 7	Prod 8	Prod 9	Prod 10	Prod 11	Prod 12	Prod 13	Prod 14
1	N	N	N	N	N	N	N	N	N	F	F	F	F	F
2	C	N	N	N	N	N	N	N	N	F	F	F	F	F
3	C	C	N	N	N	N	N	N	N	F	F	F	F	F
4	C	C	C	N	N	N	N	N	N	F	F	F	F	F
5	C	C	C	C	N	N	N	N	N	F	F	F	F	F
6	C	C	C	C	C	N	N	N	N	F	F	F	F	F
7	C	C	C	C	C	C	N	N	N	F	F	F	F	F
8	C	C	C	C	C	C	C	N	N	F	F	F	F	F
9	C	C	C	C	C	C	C	C	N	F	F	F	F	F
10	C	C	C	C	C	C	C	C	C	F	F	F	F	F
11	C	C	C	C	C	C	C	C	C	C	F	F	F	F
12	C	C	C	C	C	C	C	C	C	C	C	F	F	F
13	C	C	C	C	C	C	C	C	C	C	C	C	F	F
14	C	C	C	C	C	C	C	C	C	C	C	C	C	F
15	C	C	C	C	C	C	C	C	C	C	C	C	C	C

non-clocked products

clocked products

So, these 15 scenarios with an increasing number of clocked products are the basis for the influence test of the mixed integer resource allocation. For this all scenarios were scheduled

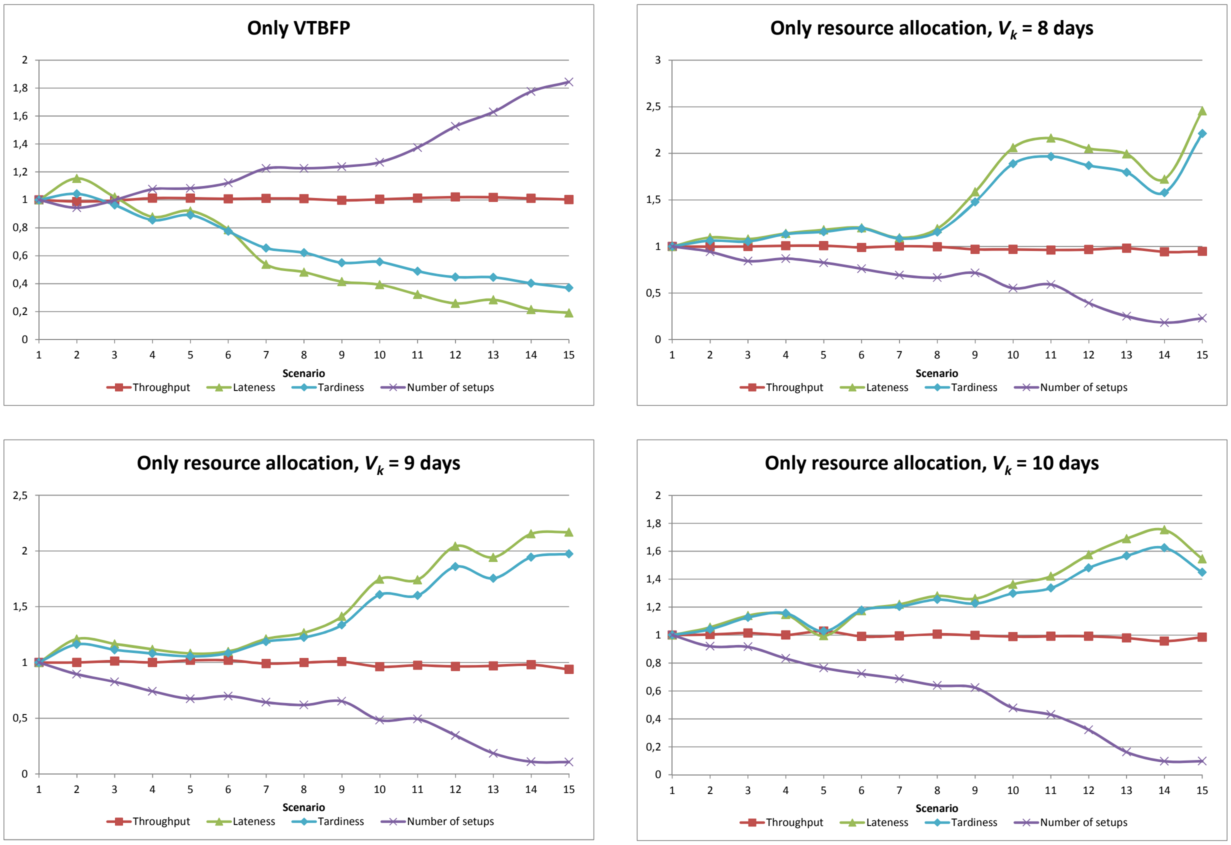
1. without a resource allocation and with VTBFP scheduling,
2. with resource allocation and without VTBFP scheduling,
3. with a resource allocation and VTBFP scheduling.

Further tests are done with the availability of the machines, so the resulting dedication matrix for each scenario have different dense. The resource allocation method is executed with the optimization model 1 for each scenario and different availabilities V_k, where the availability differs between 8 and 10 days. Thereby the used MIP solver is the IBM ILOG CPLEX solver (IBM ILOG CPLEX optimizer 2012) and at maximum 60 seconds per problem were allowed.

4.4 Results

First, the effect of the two combined methods – the VTBFP and the resource allocation – is investigated separately on the wafer test. Than the combined method is considered. The determined objectives are the throughput, the tardiness, the lateness and the number of setups. All results in the figures in this section are normalized to the value of scenario 1. This scenario has in all investigations the same properties. It is completely without using the VTBFP scheduling method and without using a resource allocation, which means, the original non-reduced dedication matrix is used. So the y-axis represents the changes in the scenarios in contrast to scenario 1, which is always normalized to 1.

In Figure 7 the results for the isolated methods VTBFP and the resource allocation are shown for each scenario. Thereby, the upper left figure shows the result from the scheduling with the VTBFP method without the upstream capacity planning. The graphic shows that the lateness and tardiness gets much better, but the number of setups increases to nearly 200%. The other three graphics in Figure 7 describe the results for the resource allocation method without VTBFP scheduling for different availabilities V_k. These three diagrams have nearly the same result: The lateness and tardiness increase significantly, whereby the number of setups decreases rapidly. In all graphics the throughput keeps nearly constant.


Only VTBFP

Only resource allocation, V_k = 8 days

Only resource allocation, V_k = 9 days

Only resource allocation, V_k = 10 days

Figure 7: Results for the separated methods VTBFP and resource allocation (the y-axis represents the changes in the scenarios in contrast to scenario 1, which is always normalized to 1)

The combination of both methods – the VTBFP and the resource allocation method is shown in Figure 8. Here only the result for the availability V_k = 9 days is presented. The other availabilities generates similar results. This diagram shows that the combination of both methods results in an improvement for all used objectives if only part of the products is clocked.

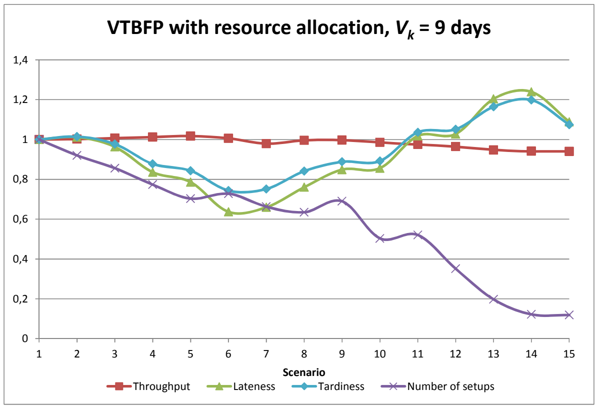


Figure 8: Results for the VTBFP method combined with resource allocation method

5 CONCLUSION AND OUTLOOK

The resource allocation is normally not suitable for the scheduling of manufacturing processes, but it is a possible method to help scheduling systems improve utilization of resources. The effect of a capacity planning is tested on a sub problem in the wafer test and there it reaches good results by reducing the makespan of schedules generated with a DES system significantly (about 9%). Simultaneously, the number of setups could be reduced by 50% with the generated test instances and the same setup strategy respectively dispatching rule. This result is not directly transferable to reality, therefore a modification of the capacity planning was used to support a clocked scheduling method – the Virtual Time Based Flow Principle – for planning the wafer test. Here the effects of both combined methods (VTBFP and capacity planning) were investigated separately. Thus, it could be figured out that the VTBFP method reduce the tardiness, but increase the number of setups. On the other hand the capacity planning has the effect on a significantly reduced number of setups while the tardiness increases. This can be explained by the reduced dedication matrix, which is a result of the capacity planning. By reducing the number of allowed machines per product, the machines do not have to change the products so often. However, the equipment allocations are also reduced for the scheduling system and therefore capacity can get lost. But the combination of both methods show good results in all investigated objectives if only a part of the products is clocked. So it could be shown, that the capacity planning is an important utility to help the scheduling with VTBFP getting good results.

Further research has to be done in other areas of the semiconductor industry. For example a capacity planning model for reticle scheduling in the lithography step seems to be promising.

ACKNOWLEDGMENTS

This work was supported by the Federal Ministry of Education and Research of Germany (promotion number 13N11588).

REFERENCES

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

Doleschal, D., J. Lange, and G. Weigert. 2012. "Mixed-integer-based capacity planning improves the cycle time in a multistage scheduling system." In *Proceedings of the 22th International Conference on Flexible Automation and Intelligent Manufacturing*.

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*, 1862-1870.

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*, 1741-1748.

IBM ILOG CPLEX optimizer. 2012. “IBM - Mathematical Programming: Linear Programming, Mixed-Integer Programming and Quadratic Programming - IBM ILOG CPLEX Optimizer - Software.” Accessed May 23. http://www.ibm.com/software/integration/optimization/cplex-optimizer/.

Keil, S., G. Schneider, D. Eberts, K. Wilhelm, I. Gestring, R. Lasch, and A. Deutschlander. 2011. “Establishing continuous flow manufacturing in a Wafertest-environment via value stream design.” In *Advanced Semiconductor Manufacturing Conference (ASMC), 2011 22nd Annual IEEE/SEMI*, 1-7.

Klemmt, A., and G. Weigert. 2011. “An optimization approach for parallel machine problems with dedication constraints: Combining simulation and capacity planning.” In *Proceedings of the 2011 Winter Simulation Conference*, 1986-1998.

Lange, J., G. Weigert, S. Keil, D. Eberts and R. Lasch. 2012. “Introducing the Virtual Time Based Flow Principle in a high-mix low-volume wafer test facility and exploring the behavior of its key performance indicators.” In *Proceedings of the 2012 Winter Simulation Conference.*

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

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*, 1857-1861.

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*, 727-735.

AUTHOR BIOGRAPHIES

DIRK DOLESCHAL studied mathematics at Dresden University of Technology, Germany. He obtained his degree in 2010 in the field of optimization. He has been a Research Assistant at Electronics Packaging Laboratory of the Dresden University of Technology since 2010 and works on the field of production control, simulation & optimization of manufacturing processes. His email is doleschal@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.

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. He was involved in development of simulation systems as well as in its application in industrial projects for scheduling. His email is Gerald.Weigert@tu-dresden.de.

ANDREAS KLEMMT received his master’s degree in mathematics (2005) and Ph.D. in electrical engineering (2011) at the Dresden University of Technology. He works as an operations research and engineering expert at Infineon Dresden. His current research interests are capacity planning, production control, simulation & optimization. His email is Andreas.Klemmt@infineon.com.