Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

MODELLING AND MATHEMATICAL OPTIMIZATION FOR CAPACITY PLANNING OF A SEMICONDUCTOR WAFER TEST MODULE

Julia Siess

University of Regensburg
Universitätsstraße 31
93049 Regensburg, GERMANY

Hermann Gold

Infineon Technologies AG
Wernerwerkstrasse 2
93053 Regensburg, GERMANY

Thomas Ponsignon

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

ABSTRACT

This paper focuses on scheduling and capacity planning problems in semiconductor wafer test. The planning of wafer test as the final stage of the semiconductor frontend manufacturing process is very complex due to many uncertain factors. Yet the processes and system allocation are not fully automated at Infineon Technologies AG in Regensburg, but partly involved with manual handling. For this reason, a mathematical optimization program has been developed to compute a realistic delivery plan including a machine allocation plan. Various data on internal resources and dedication matrices were collected by specialized departments and serve as a basis for optimization, which was carried out with IBM OPL CPLEX Optimization studio using mixed-integer programming. The focus was on the optimization of delivery due date at the lowest possible costs in terms of setup and capacity.

1 INTRODUCTION

With a turnover of over 468 billion US-dollar in the year 2018, the semiconductor industry grows steadily and is one of the key sectors for further electronic equipment (Holst 2020). Chip production always starts with a silicon wafer that is processed in the frontend. After all layers on the chip have been made, a wafer is ready to be sawn into individual chips, but before that, it must be tested to ensure that all functionalities work flawlessly. These functional tests are made in the so called “wafer test area”, which is located immediately before the exit point of a semiconductor frontend manufacturing facility (called fab in the sequel). The wafer test area consists of diverse measuring technologies under extreme temperatures. Furthermore, every lot needs to undergo several measurements which are sequenced in a different order according to the product type. Due to seasonality and other external factors from market, the demand quantity and product-mix change continuously. Usually, also the time of entry of a particular lot of a product type and its due date are typically uncertain in the semiconductor industry. The waiting times are difficult to predict and in case of not predictable incidents, they can vary by days (Uzsoy et al. 1992a, 1994).

In our problem setting, we used a fixed internal arrival forecast without any variations to set up a mathematical model. The main objective of our model is to reduce the high waiting and setup times per lot and machine, which would result in more productive uptime and a significant cost saving potential for the production area under consideration.

978-1-7281-9499-8/20/$31.00 ©2020 IEEE

2 LITERATURE REVIEW

In the literature, selected journals and encyclopedias like Dabrowiecki and Bucher (2016), Sautter and Weinerth (1993) or Aitken and Maxwell (2000) provide good technical descriptions of the technology and processes that are essential to understand and set up the mathematical model in the wafer test area.

In general, scheduling and planning in the semiconductor production and test area is a frequently and long discussed topic (Uzsoy et al. 1992a, 1992b, 1994). Nevertheless, there is still no satisfactory model that covers the entire complexity of scheduling and planning in semiconductor manufacturing. Most approaches to this topic are based on heuristics or mixed-integer programming (MIP). Especially MIP models like the model of Klemmt and Mönch (2012) for flow shop scheduling problems with time constraints or the detailed formulations and problem descriptions for fabs of Klemmt (2012) can be used as a basis for further development of MIP models. Maleck et al. (2017) provides a MIP model with time constraints and analysis risk parameters for tool interruptions in the semiconductor production area and shows an adequate method to verify the quality and robustness of the model. This verification of the model is very important to show whether it is suitable and applicable for reality. But the focus of this work should not be on the production area and instead on testing. Only a few people have dealt with this special topic so far. Doleschal et al. (2012) provides a flow line scheduling approach using a MIP model for resource allocation and later a discrete-event simulation for feasible schedules in a wafer test facility. This separation in two or more stages of optimization is considered not suitable for our model setting, and thus we focus only on one single optimization run.

For the optimization in one step, Lu (1997) uses a model with the graph theory for the assignment of the jobs to the testers. This applicable, but not yet optimal model was then further developed in Ellis et al. (2004), which shows a heuristics to solve the problem within the wafer test and even includes the different temperatures of the testers. Complicating factors for an optimization model are sequence-depending setup times and multiple criteria, which are analyzed by Huang and Lin (1998), which proposes an `interactive computer aided scheduling system' (ICASS) due to these two difficulties.

In addition, the wafer test area consists of different tester groups with different measuring types and are passed through by a lot in a specified order. These groups can be classified in so-called closed machine sets (CMS), as suggested by Gold (2012), allowing them to be viewed individually, thus reducing complexity.

3 MODELLING ASSUMPTIONS

To develop an appropriate model, we first had to define some assumptions to keep our model simple but still realistic. The following assumption were developed and set up according to own experiences and experts of the wafer test.

- Firstly, no company can process wafers without the well-known production factors and resources, which include human work, machines, financial capital or knowhow. The persons, who work in the wafer test, are called operators and are one of the most limiting and uncertain factors. For the following model, we assume that experts operating and handling the testers are always sufficiently skilled and sufficient in number. To ensure a high-end 24 hours/day production in 365 days/year, these experts work in a shift-system. Between two shifts, the operators have a short handover and discussion with the previous shift. During this time, a downtime of 10 to 30 minutes can be estimated and this three times a day.
- For testing a model an appropriate data set should be available. To gain data and especially realistic process or measuring times, we used historical data sets. On this basis we counted the median of every measuring time per wafer of every product. But in addition to the median we also assume that we have standardized lots with 25 wafers. We know that there is often the case of lots with less than 25 wafers, but this reduced wafer number is compensated by single re-measurements.
- Furthermore, a typical company needs to deal with many external unpredictable factors like power failure or gas leaks and after these events the production needs to be ramped up slowly. Therefore, additional time needs to be estimated.

- In addition, the machines and testers in the fabs cannot always be used for productive lots. Instead, development lots are tested and maintenance service works are made. These two cases and especially developments have priority which means that productive lots must even be stopped and processed afterwards.
- As granularity for the time unit days were chosen. The minutes per day represent the capacity for productive lot testing per day. A portion of the tester time per day is reserved for development and maintenance. Cutting the time unit by half would yield a more detailed and more realistic schedule at the cost of an increase of computing time roughly by a factor of four.
- Generally, in the semiconductor industry several customers or delivery models exist. According to these circumstances the planning department uses a typical pull-strategy to schedule lots. This strategy was also adopted for the developed model, which means, the lots can be tested in advance, but they need to be finished at least at the demand date.
- To implement and measure a model, it is better to use a small data set which means in our case to use one machine group instead of the whole wafer test at one side. Therefore, we formed groups of testers or used existing groups. Also, in the literature this is a problem discussed and our machine groups are so called Closed Machine Set (CMS) (Gold 2012). This approach allows all possible parallelization, but limitations due to durables are taken into account at the same time. Furthermore, care was taken to build the model and problem setting as close as possible to the real fab situation in Regensburg where testing equipment is disjunctive, thus qualification was never a problem.
- The dependency of process times of probe card has not been focused in this work. Still the mathematical model could easily be enhanced to cover probe card dependent process times.
- Since the optimization algorithm developed further on is quite complex and the requirements on computational efficiency of the real world application envisioned are very high, we use a decomposition approach to reduce the number of constraints and variables to be expected in the optimization instances typically appearing in our testing area. In a similar vein as described in Gold (2012), a given category of functional tests is called a job class when all tests belonging to this category are released on one and the same set of testers and have one and the same vector of test times in common on the so-defined set of testers. Now, we define CMS which are supposed to delimit the subsystem to be considered for any given optimization instance. A CMS consists of a minimal set testers from which neither load can be shifted from its inside to its outside nor in the reverse direction. Formally two testers x and y are in the same CMS if and only if there is a chain of job classes $z_1, \dots, z_i$ and a set of machines $m_1, \dots, m_{n-1}$ such that z_1 can be processed on x and m_1, z_i can be processed on m_{i-1} and m_i, for $i = 2, \dots, n-1$, and z_n can be processed on m_{n-1} and y. In effect, the CMS can be considered as a partitioning of the tester area into aggregate sets of job classes and related sets of machines behaving independent from each other. The algorithmic generation of all CMS is achieved via a graph theoretical concept. A graph $G(V, E)$ is defined as follows. Each tester is associated with vertex $v \in V$ of the graph and an edge is drawn between any two vertices v_i and v_j if there exists a job class which is released on the tester associated with vertex v_i as well as on the tester associated with vertex v_j. The partitioning of the tester area into CMS is established by generating the connected components of the graph $G(V, E)$.

In summary, according to these assumptions, only an uptime for productive lots of 85% is estimated and all time components are illustrated in Figure 1.

4 MATHEMATICAL MODEL

In the developed model, the different products get an index with their product number and in a similar vein, all machines are numbered with their machine number within their CMS. Auxiliary resources are the probe cards, which are product specific and have to be replaced in case of a product change. A typical time horizon for a model has always 21 days, which means that 21 time periods or time slots are within an optimization run. These four indices are listed in the Table 1. These indices are also essential for the decision variables, which are calculated during the optimization run by CPLEX and shown in Table 2. The most important decision variable is x, which represents the production quantity of every product on a certain machine in a certain time slot. Along with this, an indicator variable *y* is introduced, which indicates the specifics of the resource consumption, namely the tester *m* used during a particular time slot *t* for the production of a certain product *p*. The use of the indices *t*, *p*, and *m* is deployed also for other variables having the same dimensions as *y* in the sequel.

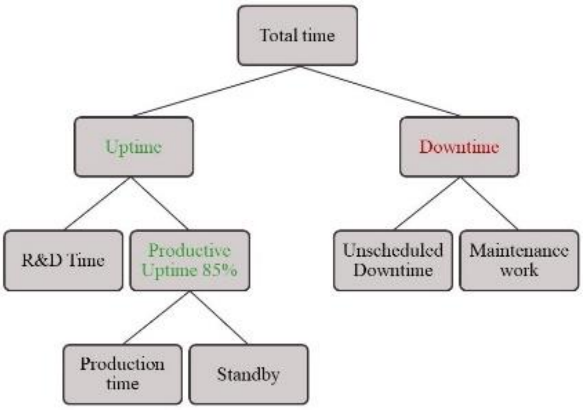

Figure 1: Own illustration based on Pomorski (2019) and Maleck et al. (2017)

Table 1: Resources

$p \in P$	Products
$m \in M$	Machines / Testers
$nk \in NK$	Probe cards
$t \in T := [Tmin, Tmax]$	Time period

Table 2: Decision variables

$x_{t,p,m} \in \mathbb{N}$	Production quantity
$s_{t,p} \in \mathbb{N}$	Stock quantity
$y_{t,p,m} \in \{0; 1\}$	Production decision
$pc_{t,p,m} \in \{0; 1\}$	Setup change decision
$\delta \in \{0; 1\}$	Delta for Big-M of temperature
$a_{t,p}^{+} \in \mathbb{R}$	Temperature difference for cooling
$a_{t,p}^{-} \in \mathbb{R}$	Temperature difference for heating
$s\delta \in \{0; 1\}$	Delta for Big-M of not yet tested stock / backlog
$s_{t,p}^{+} \in \mathbb{N}$	Stock of not yet tested lots
$s_{t,p}^{-} \in \mathbb{N}$	Backlog of not yet tested lots
$g \in \mathbb{R}$	Additional needed capacity

If a certain *x* is positive, then the corresponding *y* also needs to be positive. Furthermore we need to ensure that production always guarantees the delivery due date of the products. This implies that it must be possible to pre-produce products since otherwise resource conflicts during a particular time slot may cause the problem to become infeasible. Pre-produced products are put into stock, the associated variable denoted by *s*. Satisfaction of demand from stock is considered to occur without delay. Like *y* also *pc* is binary and represents the product change, which means that the setup and probe card on a machine are changed. But after a setup change also the product specific measuring temperature needs to be modified. Therefore, the Big-M method, which is a method for solving linear programming problems, was integrated in the model. This mathematical method contains a Big-M that can be seen as a placeholder with a high number for a certain number, which is not defined at the beginning. This unspecific Big-M is used in the model to simulate the temperature difference and the difference of the not yet tested lots in stock, which is shown in Table 3. Both differences can be positive or negative and to set it into account the amounts need to be absolute. To ensure that only one case of both, e.g. heating or cooling / stock or backlog, will occur, in both variants an indicator variable δ is added. These two aspects of the method have been implemented in the optimization model in CPLEX.

Table 3: Big-M

$BS \in \mathbb{N}$	Big-M for not yet tested stock or backlog
$B \in \mathbb{N}$	Big-M for maximal temperature difference

Especially the stock or backlog of the not yet tested lots delivers an additional value. In case of a backlog the planning team can give early warnings to the other departments and flashes lots, which means that they get a higher priority status and are transported and processed directly.

The model depicted so far may yield relaxed solutions or return a problem as to be infeasible during congestion periods. For these cases a variable reflecting necessary excess capacity c is defined. This additional capacity is considered quite expensive and weighted comparatively high in the goal function. If it is needed, the planners can decide to reschedule the demand plan, buy additional external capacity or organize additional transports to other internal fabs to shift capacity. The constraint including excess capacity guaranties that there is always a feasible solution, but it came rarely into play.

Another constraint is given by the dedication matrix *n*, which represents the possible assignments of products to machines. The created dedication matrix is quite important, because not every product is allowed to be measured on each machine. The machines have different qualifications and technologies, which is reflected in this matrix. Furthermore, the probe card dedication matrix has to be taken into regard when assigning lots to testers. The number of testers allocated to some given product p during any particular time slot can never exceed the number of probe cards available for this product p, which latter is given by the row sum of row p in matrix $nkm_{p,m}$.

During one time slot the lots of the pre-operations arrive in arbitrary order and are called pre-products or not yet tested lots, because they are one step back in the workflow. The forecasted vector for these arrivals is expressed by v and in opposite to it, d represents the demand call.

All of above parameters listed in Table 4 have two dimensions. The product number, the initial stock, the setup time per product change, the setup time per lot start, the cycle or measuring time and the measuring temperature all depend only on product type. Differently to them, the cost rates are generalized and can be seen in Table 5.

Table 4: Fixed input factors

$c_{t,m} \in \mathbb{R}$	Capacity
$n_{p,m} \in \{0, 1\}$	Dedication matrix for products and machines
$nkm_{p,m} \in \mathbb{N}$	Probe cards dedication matrix
$v_{t,p} \in \mathbb{N}$	Arrivals of not yet tested lots
$d_{t,p} \in \mathbb{N}$	Demand

Table 5: One-dimensional input factors and cost rates

$pn_p \in \mathbb{N}$	Product number	$qc \in \mathbb{R}$	Setup cost rate
$sini_p \in \mathbb{N}$	Initial stock of products	$h \in \mathbb{R}$	Stock cost rate
$qp_p \in \mathbb{R}$	Setup time per product change	$svc \in \mathbb{R}$	Stock cost rate for not yet tested lots
$ql_p \in \mathbb{R}$	Setup time per lot start	$blc \in \mathbb{R}$	Backlog cost rate for not yet tested lots products
$lt_p \in \mathbb{R}$	Cycle time per product	$tc \in \mathbb{R}$	Temperature cost rate
$f_p \in \mathbb{R}$	Test temperature per product	$gc \in \mathbb{R}$	Additional capacity cost rate

After defining all necessary variables, we now establish formulas relating these variables to each other. Below eight decision expressions are listed, which constitute the comprehensive objective function.

Breakdown of the objective function in decision expressions:

$$\text{Product setup costs} = \sum_{t=1}^{Tmax} \sum_{p \in P} \sum_{m \in M} (pc_{t,p,m} * qp_p * qc) \tag{1}$$

$$\text{Stock costs} = \sum_{t=Tmin-1}^{Tmax} \sum_{p \in P} \sum_{m \in M} (s_{t,p} * h) \tag{2}$$

$$\text{Backlog costs} = \sum_{t=1}^{Tmax} \sum_{p \in P} \left(s_{t,p}^{-} * blc \right) \tag{3}$$

$$\text{Cooling costs} = \sum_{t=1}^{Tmax} \sum_{m \in M} \left(a_{t,p}^{+} * 2 * tc \right) \tag{4}$$

$$\text{Heating costs} = \sum_{t=1}^{Tmax} \sum_{m \in M} \left(a_{t,p}^{-} * tc \right) \tag{5}$$

$$\text{Additional capacity costs} = \mathrm{g} * \mathrm{gc} \tag{6}$$

Objective function:

Minimize Product setup costs + Stock costs + Backlog costs +
Cooling costs + Heating costs + Additional capacity costs (7)

5 CONSTRAINTS

In the following, all auxiliary conditions are mentioned and briefly explained. The scope of validity for the index sets used in this sequel are defined as follows: $\forall t \in Tmin..Tmax, p \in P, m \in M.$ To optimize the problem within the given framework, several constraints were set up.

One of the most important restrictions is the capacity limitation, because only the available uptime can be used.

$$c_{t,m} + g \geq \sum_{p \in P} (x_{t,p,m} * (lt_p + ql_p)) \tag{8}$$

Furthermore the current stock of the planning horizon need to be counted new in every time slot to decide if an upcoming demand will be fulfilled by the current stock or with a new production quantity.

$$s_{t,p} = s_{t-1,p} + \sum_{m \in M} x_{t,p,m} - d_{t,p} \tag{9}$$

In addition, there can be an initial stock at the beginning of a planning horizon, that is described as the stock in period zero.

$$s_{Tmin-1,p} = sini_p \tag{10}$$

To save stock cost also the final stock of a planning horizon should be zero, otherwise too much quantity would be produced.

$$s_{Tmax,p} = 0 \tag{11}$$

The planning department follows the pull-strategy which is displayed by the local and global “Variable Upper Bound” or VUB. This construct is a special constraint according to Pochet and Wolsey (2006). Formula 16 represents the local VUB and formula 17 the global VUB. The local VUB ensures that the pull-principal is fulfilled which means that *x* only can be as high as the result of the multiplication of *y* and the sum of the total demand of the remaining periods. This case to avoid an overproduction would be already excluded with the last formula above, that the final stock is zero. But the VUB also ensures that the production brings up cost. Furthermore, the local VUB had to be extended to a global one because a CMS consists of more than one machine. For this case the production quantity *x* is summarized over all machines, but the same for *y* is not possible! If *y* also is summarized and e.g. three machines are in use, then the sum of all *y* during one time slot would be three and the remaining demand would be multiplied with three. This would result, that *x* could be three times higher than it should be. To prevent this case a minimum function was added to the VUB, that only the minimum of the sum of *y* or 1 is used. If the summarization is zero than the right hand of the global VUB is zero and otherwise the demand is always multiplied with one instead of two or another higher number.

$$x_{t,p,m} \leq y_{t,p,m} * \sum_{z=t}^{Tmax}(d_{z,p}) \quad (12)$$

$$\sum_{m \in M}(x_{t,p,m}) \leq \min\left(\sum_{m \in M}(y_{t,p,m}), 1\right) * \sum_{z=t}^{Tmax}(d_{z,p}) \quad (13)$$

Furthermore, not every product can be tested on every machine within one CMS due to different technologies or qualifications.

$$n_{p,m} \geq y_{t,p,m} \quad (14)$$

This dedication problem also applies to the probe cards. The probe cards are product specific and only available in a certain number.

$$\sum_{nk \in NK} nkm_{p,nk} \geq \sum_{m \in M} y_{t,p,m} \quad (15)$$

Another problem, which is an essential assumption of the model, is that only one product can be tested during one time slot and a setup product change is only possible at the beginning of a new period.

$$\sum_{p \in P} y_{t,p,m} \leq 1 \quad (16)$$

As it could be seen for the initial stock, we implemented a period zero for the initialization. During this period nothing will be produced.

$$y_{Tmin-1,p,m} = 0 \quad (17)$$

In addition, *y* is not allowed to be bigger than *x* and it is also binary, but if there is no production quantity within a period *(x = 0)*, then *y* also needs to be zero.

$$x_{t,p,m} \geq y_{t,p,m} \quad (18)$$

A very important constraint is also the product change which needs to be simulated for all different cases. In this model only a new setup is been charged with costs, because during an idle period the last setup stays and is not removed.

$$y_{t,p,m} - y_{t-1,p,m} \leq pc_{t,p,m} \quad (19)$$

Furthermore, we had to simulate the variation of the not yet tested lots in stock, which can rise or decrease. To gain a high benefit of this we even allow a backlog in the model, which means, the model needs to deliver by the due date, but it is able to do pre-production without any constraint violation. Therefore, we used the Big-M method combined with absolute values like Kallrath (2013) to simulate the changes of all not yet tested products.

$$\sum_{w=Tmin}^{t}(v_{w-1,p}) - \sum_{w=Tmin}^{t} \sum_{m \in M}(x_{w,p,m}) = s_{t,p}^{+} - s_{t,p}^{-} \quad (20)$$

$$s_{t,p}^{+} \leq BS * s\delta_{t,p} \quad (21)$$

$$s_{t,p}^{-} \leq BS * (1 - s\delta_{t,p}) \quad (22)$$

At first the sum of the arrivals of the not yet tested lots is reduced by the test quantity of the related product and their difference is represented by $s_{t,p}^{+}$ and $s_{t,p}^{-}$. In detail this means that $s_{t,p}^{+}$ represents a positive stock of the not yet tested lots and $s_{t,p}^{-}$ a backlog of them. In addition, only one of both is allowed to have a value while the other one needs to be zero. This is achieved by two additional constraints where the binary variable $s\delta_{t,p}$ or $(1 - s\delta_{t,p})$ is multiplied with BS which represents the highest possible stock or backlog of the not yet tested lots. Especially the information of a possible backlog is an additional value for the planners because as a result of it they can give early warnings to the other planning departments and they can react in advance.

Another important part of the model is the optimization of the reduction of the setup time which also heavily depends on the heating or cooling time of a machine for its next test. Regarding the reduction of these setup times the temperature difference should be as low as possible. To achieve only small temperature changes the heating or the cooling difference need to be measured and penalized in the goal function. Therefore, the temperature changes are also simulated similar with the Big-M method like the not yet tested lots in stock discussed in the last section.

$$\sum_{p \in P}(y_{t-1,p,m} * f_p - y_{t,p,m} * f_p) = a^+_{t,m} - a^-_{t,m} \tag{23}$$
$$a^+_{t,p} \leq B * \delta_{t,p} \tag{24}$$
$$a^-_{t,p} \leq B * (1 - \delta_{t,p}) \tag{25}$$

The heating or cooling of one machine is represented by $a^+_{t,m}$ or $a^-_{t,m}$ and only one of them can be positive and the other one needs to be zero.

Finally, also the non-negative conditions need to be added to ensure that negative values are not assigned to any decision variable. In the optimization program these constraints do not need to be written down, because the variables are already defined as positive integer variables.

6 APPLICATION AND IMPLEMENTATION FOR THE CASE OF WAFER TEST AT INFINEON TECHNOLOGIES AG IN REGENSBURG

For implementing the model, we used IBM ILOG CPLEX Optimization Studio and an Excel file for the data sets. In summary, we run the model with different tester groups as well as historical data. At first, we collected data from all testers, probe cards, products and internal forecasts to set up a database and build matrices. Then, we implemented the data in Excel files with the right format so that CPLEX could directly read the data from the prepared files. In addition we wrote the model itself in a mod.-file and the instructions with all reading and writing commands in a dat.-file. After that a run configuration was created to run the model. Depending on the complexity, the machine and product quantity, CPLEX needed a few seconds or up to one minute. Due to the regarded time slot with the mix of products and their quantity also the computing time and the optimization potential differ. To vary the results and see the stability also different time slots where chosen, but for comparison the same slot was taken, for instance, a performance comparison was made. Therefore, the forecasted data set and the historical data set of a certain CMS was analyzed. All quantities, which were measured in the past, result in demands to ensure the same quantities of the respective products and the same initial situation or conditions.

In addition, we assume that delivery of a given quantity of tested lots is promised and that the corresponding lots, once they are tested, feature the presumed yield level. In case this target yield level is not sufficiently met for a given product, corresponding stock levels in the supply chain will be depleted earlier and trigger additional production, incoming WIP and demand in the near future.

7 RESULTS

The results of these optimizations are allocation plans for every machine group like shown in Table 6. Above one can see all 21 time slots and on the left side all machines of a CMS are listed. In addition for every machine the lot quantity and the respective product number are shown, which means, for instance in period 2 one lot of product 31 on machine 1 and two lots of product 15 on machine 2 are tested. Hence, the planning department gains an additional benefit of a plan for a horizon on which machine how many lots of which product should be tested to ensure to be in time with all deliveries.

Furthermore, a key performance indicator should be introduced to measure the performance of the optimization model, because the objective of the model is to reduce the setup costs and to test as many lots as possible of the same product in succession. Therefore, we counted the measured lots per setup. As an example on machine 1, we have 27 lots and 7 setup changes, which means we have 3.8571 lots or runs per setup on this machine. If we have an idle period between the setup of the same period we still count a new setup, because a development lot or a new setup version could be started. If we also weight it, then we use the percentage of the lots on this machine divided through the total lots of this CMS. This means the percentage is 27/274 = 9.85% and this multiplied with 3.8571 is 0.38. If we calculate it for the total CMS then we get a KPI of 5.7432 runs per setup *R/S* for it. This value can be compared with the KPI of the real historical data set with 2.29 *R/S*, which was optimized in advance with a heuristics. From this comparison, it can be seen that the value of the machine allocation plan of the optimization model is more than 2.5 times higher than the historical one.

Table 6: Optimized allocation plan

	Machines / Time	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
M01	Sum of x		1	1	1	1	1		1	2	2	1	1	1	1	1	1	2	5	2	1	1
	Product nb		31	31	31	31	31		28	19	24	24	24	24	24	28	28	15	17	17	17	17
M02	Sum of x		2	1	2	1			2	1	2	1	1	1	1	2	1	1	2	2	2	1
	Product nb		15	15	24	24			16	21	15	31	31	31	31	4	4	13	13	4	4	20
M03	Sum of x		1				1	1	1	1	1	1	1		1	1	1			1		1
	Product nb		32				32	32	32	27	27	27	27		33	33	33			32		26
M04	Sum of x	3	3	1	1	2			2	1	1	1	1	2	1		2	2	2	1		2
	Product nb	12	12	22	22	12			22	12	12	12	12	13	13		21	22	19	19		24
M05	Sum of x		1	2	2		1	2	1	1			4	1	1	1	1	2	2	2		
	Product nb		4	4	4		3	3	29	29			1	3	3	3	2	2	2	2		
M06	Sum of x			2	1	1	1	1	1	1	2	2	2	1	1	1	2	2	2	2	2	6
	Product nb			2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1
M07	Sum of x						1	1	1	1					1	1	1	1		1	1	
	Product nb						30	30	30	30					29	29	29	29		3	3	
M08	Sum of x	1		1		1	2	1	1	1	2	1			2	2	1	2	1	2	1	3
	Product nb	1		19		18	18	18	18	18	18	18			20	16	18	18	18	18	18	18
M09	Sum of x	1	3	1			2	1	2	2	2	2	2	1	1	1	1	1	1	1	1	2
	Product nb	5	5	5			6	11	11	11	11	11	5	7	7	7	7	7	7	7	7	11
M10	Sum of x	2	2	2			1	2	2	2	2	2	1	2	2	2	2	2	2	1	1	
	Product nb	11	11	11			25	25	25	11	11	11	11	11	11	11	11	11	11	9	9	
M11	Sum of x	2	2	2			1	1	2	2	2	1	1	1	1	1	2	2	1	1	1	2
	Product nb	11	11	11			7	7	11	11	11	9	9	9	9	9	11	11	25	25	25	25

Table 7: Summary of comparison results

Comparison	**R/S**	**Total Setups**
Historical allocation	2.29	133
Optimized allocation	5.74	67

This result also shows a high saving potential of setup time or a potential of additional productive uptime. In the historical data set we had 133 setups instead of 66 in the optimized version, which are 67 setup changes less. If we count with the minimum setup time of 10 minutes, we have 670 minutes or over 11 hours more of setup time in the historical and real data set within one planning horizon of 21 days. If we extrapolate this example for one year, we get a saving potential of over 8 days or 9.5 days of an uptime with 85%. To be realistic most of the operators need more than 10 minutes for a setup with a product change especially for a complex product where they need to change the probe card. Moreover, they need about 15 minutes in average. If we count the less 67 setup changes with a changing time of 15 minutes, than we get 1005 minutes or 16.75 hours in a planning horizon for 21 days. If we extrapolate again with 15 minutes setup change time for one year, then we get a time saving potential of over 12 days with 24 hours or 14.27 days with 85% usable uptime.

In summary, we save or have more uptime of 2.61% for a setup change with a duration of 10 minutes and of 3.91% with 15 minutes. Obviously, this time saving result is a big cost saving potential. The cost saving potential for the respective CMS is estimated at a six-digit euro range. If we set all CMS only in one location into account, then we estimate a seven to eight-digit euro range.

Furthermore, we determined connections when we started a sensitivity analysis. We changed only the setup cost rate to analyze the reactions of the model to ensure if the model is suitable and predictable for reality. During this sensitivity analysis we investigated the reactions of the value of the objective function, the quantity of the iterations, *R/S*, and the computing time in average. The setup cost rate *qc* is directly in the goal function for the calculation of the product and lot setup costs, which is the reason why the target value rise linear with the setup cost rate. Then the quantity of the iterations also rises and later falls. Also the computing time rises, but not as linear as the target value. Depending on *qc*, the pure computing time varied between 0.75 and 16.5 seconds and the total process time per run in CPLEX was between 3.9 and 19.8 seconds. The KPI *R/S* increases continuously, stronger at the beginning and then decreases over the course. In summary the model delivered overall suitable and predictable results, which is very important for the model itself and for its potential usage in reality.

8 CONCLUSION

The focused problem of capacity planning is a well-known problem in manufacturing areas. Especially in the semiconductor industry it is often focused, because there is significant cost and time saving potentials, which was also the purpose for this research and also highlights the practical importance of the problem. Currently, only heuristic solutions are used in practice due to the high complexity, the difficulty of correct mathematical representation and simulation of reality. But with the help of simplifying assumptions, a MIP model could be implemented in CPLEX and could show, that it can develop substantial cost and uptime saving potentials. These results can already be used to optimize lot scheduling over horizon. But for sure the results will not be optimal and can even be higher without all mathematical assumptions.

ACKNOWLEDGEMENT

We would like to thank Hans Ehm from the Supply Chain Innovation department at Infineon Technologies AG who made the Master's thesis possible and for initiating the idea of this paper**.**

REFERENCE

Aitken, R. C., and P. C. Maxwell. 2000. "Chip Testing". In *Handbook of Semiconductor Manufacturing Technology*, edited by Y. Nishi and R. Doering, 983–997. Boca Raton: CRC Press.

Dabrowiecki, K., and J. Bucher. 2016. "Der Kontaktwiderstand beim Testen mit Wafer-Prüfkarten". *electronic fab* 10(4):9–11.

Doleschal, D., J. Lange, G. Weigert, and A. Klemmt. 2012. "Improving Flow Line Scheduling by Upstream Mixed Integer Resource Allocation in a Wafer Test Facility". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 2037–2048. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ellis, K. P., Y. Lu, and E. K. Bish. 2004. "Scheduling of Wafer Test Processes in Semiconductor Manufacturing". *International Journal of Production Research* 42(2): 215–242.

Gold, H. 2012. "Why Our Company Uses Programming Languages for Mathematical Modeling and Optimization". In *Algebraic Modeling Systems*, edited by J. Kallrath, 161–169. Berlin, Heidelberg: Springer.

Holst, A. 2020. Semiconductor market size worldwide from 1987 to 2020. https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/, accessed 11th February 2020.

Huang, S. C., and J. T. Lin. 1998. "An Interactive Scheduler for a Wafer Probe Centre in Semiconductor Manufacturing". *International Journal of Production Research* 36(7): 1883–1900.

Kallrath, J. 2013. *Gemischt-ganzzahlige Optimierung: Modellierung in der Praxis.* 2nd ed. Wiesbaden: Springer.

Klemmt, A. 2012. *Ablaufplanung in der Halbleiter-und Elektronikproduktion: Hybride Optimierungsverfahren und Dekompositionstechniken.* Wiesbaden: Springer.

Klemmt, A., and L. Mönch. 2012. "Scheduling Jobs with Time Constraints Between Consecutive Process Steps in Semiconductor Manufacturing". In *Proceedings of the 2012 Winter Simulation Conference,* edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 2173–2182. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lu, Y. 1997. *Scheduling of Wafer Test Processes in Semiconductor Manufacturing.* Ph. D thesis, Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia. https://vtechworks.lib.vt.edu/handle/10919/10153, accessed 3rd August 2020.

Maleck, C., G. Weigert, D. Pabst, and M. Stehli. 2017. "Robustness Analysis of an MIP for Production Areas with Time Constraints and Tool Interruptions in Semiconductor Manufacturing". In *Proceedings of the 2017 Winter Simulation Conference,* edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 3174–3725. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Pochet, Y. and L. A. Wolsey. 2006. *Production Planning by Mixed Integer Programming*. New York: Springer.

Pomorski, T. 2019. Major Revision Update for SEMI E10 Specification for Definition and Measurement of Equipment Reliability, Availability, and Maintainability (RAM). https://www.semi.org/en/Standards/CTR_031244, accessed 3rd August 2020.

Sautter, D. and H. Weinerth. 1993. *Lexikon Elektronik und Mikroelektronik*. 2nd ed. Berlin, Heidelberg: Springer.

Uzsoy, R., C. Y. Lee, and L. A. Martin-Vega. 1992a. “Scheduling Semiconductor Test Operations: Minimizing Maximum Lateness and Number of Tardy Jobs on a Single Machine”. *Naval Research Logistics (NRL)* 39(3):369–388.

Uzsoy, R., C. Y. Lee, and L. A. Martin-Vega. 1992b. “A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part I: System Characteristics, Performance Evaluation and Production Planning”. *IIE Transactions* 24(4):47–60.

Uzsoy, R., C. Y. Lee, and L. A. Martin-Vega. 1994. “A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part II: Shop-Floor Control”. *IIE Transactions* 26(5):44–55.

AUTHOR BIOGRAPHIES

JULIA SIESS works as a specialist for Supply Chain and Procurement in the area of semiconductor and electronics. She obtained her Master of Science degree in Business Administration with focus on Industrial Management and Informatics from University of Regensburg. Her research interests include logistics, supply chain and production planning to improve and optimize scheduling and costs. Her email address is siessjc@gmail.com.

HERMANN GOLD is a Senior Staff Engineer at Infineon Technologies AG, where he is working on planning and scheduling problems in semiconductor manufacturing. He studied computer science at the University of Erlangen and received a doctorate degree from the Faculty of Mathematics at the University of Würzburg. His special research interest is in the combination of queueing theory and optimization. His email address is hermann.gold@infineon.com.

THOMAS PONSIGNON works as a Senior Staff Engineer in the Corporate Supply Chain organization of Infineon Technologies AG in Munich, Germany. He obtained master’s degrees in Industrial Engineering from the EPF-Ecole d’Ingénieurs, Sceaux, France and the University of Applied Sciences, Munich, Germany and a Ph.D. in Mathematics and Computer Science from the University of Hagen, Germany. His research interests include production planning and simulation of semiconductor supply chains. His email address is thomas.ponsignon@infineon.com.