Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

SIMULATION-BASED OPTIMIZIATION TO DESIGN EQUIPMENT HEALTH-AWARE DISPATCHING RULES

Lorenz Reinhardt
Lars Mönch

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

ABSTRACT

In this paper, we discuss the construction of dispatching rules for semiconductor wafer fabrication facilities (wafer fabs) that take equipment health issues into account. Monitoring the equipment health status of critical machines is important to maintain process quality and to reduce rework and scrap. Usually, there is a tradeoff between quality- and productivity-related goals in wafer fabs. Obtaining an appropriate compromise between these two objectives is addressed by considering blended dispatching rules. Simulation-based optimization based on variable neighborhood search (VNS) using a reduced simulation model of a wafer fab is applied to determine appropriate weights for the different priority indices. We demonstrate by simulation experiments that the obtained blended dispatching rule performs well compared to dispatching rules that are designed only for a single quality- or productivity-related objective.

1 INTRODUCTION

Wafer fabs belong to the most complex manufacturing systems that currently exist (cf. Mönch et al. 2013). Several hundreds often very expensive machines that are organized in machine groups are used to process wafers. A wafer is a thin silicon disc. Up to one thousand chips can be produced on a single wafer by manufacturing the chips layer by layer in a wafer fab. Frequent machine breakdowns are typical for wafer fabs, i.e., the machines are highly unreliable or require preventive maintenance activities to keep them reliable. Sequence-dependent setup times occur on some machines that can be considerably longer than the processing times of the operations on these machines. Lots are the moving entities in wafer fabs. A lot can contain up to 50 wafers. Usually, several hundreds of lots can be find at the same time in a wafer fab. Reentrant flows, i.e., a lot may visit the machines of the same machine group several times commonly occur in wafer fabs. The product mix is diverse and volatile. Up to 800 process steps are required to produce chips of an advanced technology. Different process types such as batch processes and single wafer processes exist in wafer fabs. Here, a batch is a set of lots that are processed at the same time on a single machine. On-time delivery is an important performance measure in semiconductor manufacturing since the customer due dates are often very aggressive due to the fierce competition in this industry.

It is well-known that dispatching approaches are a common production control tool in many wafer fabs (cf. Sarin et al. 2011, Mönch et al. 2011). Recently, there is a need to better integrate operation control activities, for instance based on dispatching and scheduling, and advanced process control (APC) issues (cf. Yugma et al. 2015). Big data approaches allow to compute equipment health-related information more easily. However, up to now quality issues are only rarely addressed in dispatching strategies for wafer fabs (cf. Chien et al. 2015). Scheduling approaches that take into account equipment health issues are proposed by Obeid et al. (2012) and Doleschal et al. (2015). In the present paper, we apply the approach

to design blended dispatching rules proposed by Dabbas and Fowler (2003) to the situation that equipment health status-related information is available. We demonstrate that metaheuristic approaches are helpful to speed-up the simulation-based design process.

The paper is organized as follows. The problem is described and analyzed in Section 2. This includes a discussion of related work. In Section 3, we present details of the proposed dispatching scheme. Moreover, we show how VNS can be incorporated to reduce the computational burden of the simulation-based optimization procedure. The applied simulation environment is discussed in Section 4. The results of simulation experiments are shown in Section 5. Conclusions and future research directions are provided in Section 6.

2 PROBLEM DESCRIPTION AND ANALYSIS

2.1 Problem Setting

We are interested in developing a dispatching approach that meets on-time delivery performance, throughput, cycle time, and quality requirements at the same time. In the present paper, pure dispatching is preferred over a deterministic fab-wide scheduling approach since it is more flexible to deal with global performance measures and disturbances. For the sake of simplicity, the dispatching approach should be applied only to the bottleneck work center of a wafer fab.

We assume that the bottleneck machine group consists of m parallel identical machines. Each machine i has a specific eligibility for lot j with respect to quality that is expressed by a number $c_{ij} \in [0,1], i = 1,\ldots,m$. A c_{ij} value close to one refers to a large eligibility. We assume that we have a matrix $C = (c_{ij}) \in IR^{m \times n}$ where n is the number of lots to be dispatched in a certain horizon H. A procedure based on data mining to determine the entries of this matrix is proposed by Chien et al. (2015). A low eligibility of a (machine, lot) pair leads to a large amount of scrapped material and results therefore in low throughput. A eligibility value close to one refers to an appropriate health status of the corresponding machine. This means that defects produced with a machine are less likely. Moreover, we assume that we know for each lot j the due date d_j, a weight w_j that is used to represent the importance of the lot, and a release date r_j. The processing of a lot follows a product-specific route. It has n_j processing steps. The corresponding process steps are labeled by $k = 1,\ldots,n_j$, while the corresponding processing times are $p_{jk}, k = 1,\ldots,n_j$.

The different criteria considered in this paper are in conflict. For instance, a high throughput will be obtained when we always choose for each lot the machine with the highest eligibility value. But at the same time the on-time delivery performance will be low since urgent lots eventually have to be wait for a long time. A similar conflict exists for throughput and cycle time. Therefore, we are interested in designing a blended dispatching rule that allows to model preferences towards one of the specified objectives.

2.2 Related Work

We discuss related work with respect to dispatching and scheduling approaches that take into account the equipment health status of the machines. Moreover, we also briefly survey approaches to design blended dispatching rules in semiconductor manufacturing.

Chien et al. (2015) propose a statistical method to determine tool affinity to hedge the variation between the photolithography process for pattern development and the etching process to reduce the etching bias caused by tool misalignment. The obtained dissimilarity values are used to find for each lot the machine that leads to the desired etched critical dimension. This approach is used to design a dispatching rule that takes the dissimilarity value into account. However, the dispatching rule only considers the

equipment health status. Reaching specific productivity-related goals is not supported by the proposed rule.

Doleschal et al. (2015) study a mixed integer programming (MIP) formulation in a rolling horizon setting for a situation where a health value is available for each machine of a machine group. Moreover, two dispatching rules are considered within the simulation experiments. They demonstrate that the MIP approach slightly outperforms the quality-aware dispatching rule from a real-world wafer fab. But again, only quality and throughput are considered in a combined objective function.

Obeid et al. (2012) study a parallel machine scheduling problem where yield is considered as one criterion. A static and deterministic environment is assumed. The flow time and the number of disqualifications are the other two criteria of interest. The risk is modeled as the expected yield resulting from assigning a job to a specific machine. A combined objective function is considered. In addition to a MIP formulation, two yield-centric list scheduling heuristics are proposed. The importance to make integrated decisions for productivity- and quality-related objectives is also pointed out by Yugma et al. (2015).

Next, we discuss several approaches to design blended dispatching rules for wafer fabs. Dabbas and Fowler (2003) propose a two-phase procedure to construct blended dispatching rules. In a first phase, they use mixture experiments to derive a response surface-based metamodel. They then use search-based optimization methods to optimize the weights that belong to the different indices of individual dispatching rules that support different objectives. A similar approach is used by Zhang et al. (2009). In both papers, the desirability function approach by Derringer and Suich (1980) is applied to deal with the multiple-objective situation.

Li et al. (2013) and Li and Min (2016) propose blended dispatching rules where the different weights are set in a situation-dependent manner. Artificial neural networks and particle swarm optimization are used to find appropriate weight values. Only a single criterion, the number of moves, is considered. Genetic programming approaches are proposed by Pickardt et al. (2010) and Hildebrandt et al. (2014) to discover dispatching rules for wafer fabs. But only a single criterion is used. In the present paper, we extend the approach by Dabbas and Fowler (2003) to the present situation since it allows to consider several objectives at the same time. However, in contrast to Dabbas and Fowler (2003), we use simulation more directly by considering a reduced simulation model of the wafer fab. Moreover, we use a metaheuristic approach, namely VNS, to expedite the search. This allows us to avoid the response surface technology.

3 APPROACH TO DESIGN DISPATCHING RULES

3.1 Approach to Determine a Blended Priority Index

We are interested in simultaneously obtaining small cycle times, meeting the quality requirements, and striving for a large on-time delivery performance. Note that the throughput performance measure can be used as a surrogate measure for the quality-related measure since the throughput is low if a large number of lots is scrapped due to quality problems. The on-time delivery performance is measured by the total weighted tardiness (TWT) of the completed lots. It can be computed as $TWT := \sum w_j T_j$ where the tardiness of lot j is $T_j = \max(C_j - d_j, 0)$. Here, C_j is the completion time of lot j.

Each of these objectives is supported by a specific dispatching rule. It is well known that the Weighted Shortest Processing Time (WSPT) rule leads in some situations to small weighted cycle time values in wafer fabs (cf. Mönch et al. 2013). The corresponding priority index is given by:

$$I_{WSPT}(j) := w_j / p_{jl}\ , \tag{1}$$

where the process step l is performed on a machine of the bottleneck machine group. The job with the largest $I_{WSPT}(j)$ value is chosen to be processed next on an available machine. The quality-related requirements are supported by a dispatching rule with priority index:

$$I_Q(j,i) := c_{ij}\ . \tag{2}$$

Again, the lot with the largest index value is selected to be processed next on an empty machine of the machine group. Note that the index value is machine-specific.

Next, we describe a version of the Apparent Tardiness Cost (ATC) dispatching rule (cf. Vepsalainen and Morton 1989) similar to Mönch and Zimmermann (2004) that is able to take into account the global due dates. The corresponding priority index is given by:

$$I_{ATC}(j,t)=\frac{w_j}{p_{jl}}\exp\left(-\frac{\left(d_j-p_{jl}-t-\sum_{k=l+1}^{n_j}\left(wt_{jk}+p_{jk}\right)\right)^+}{\kappa\bar{p}}\right), \tag{3}$$

where t is the time for decision-making, wt_{jk} is the estimated waiting time for process step k of job j, κ is a scaling parameter, and $\bar{p}$ is the average processing time of the jobs waiting in front of the bottleneck machine group for processing. In addition, we use the abbreviation $x^+ := \max(x,0)$. Note that the quantity $\tilde{d}_j := d_j - \sum_{k=l+1}^{n_j}\left(wt_{jk}+p_{jk}\right)$ can be interpreted as a local due date with respect to the bottleneck machine group. In this research, we use straightforward waiting time estimates of the form $wt_{jk} := (FF-1)p_{jk}$, where FF is the ratio of $C_j - r_j$ and the raw processing time, i.e. the quantity $\sum_{k=1}^{n_j} p_{jk}$. Therefore, we assume that the waiting time of a certain step depends in a linear way on the processing time of this step. Again, the lot with the largest index value is chosen to be processed next. We combine the three indices (1), (2), (3) to a blended index of the form:

$$I(j,i,t,\lambda_1,\lambda_2,\lambda_3) := \lambda_1 I_Q(j,i)+\lambda_2\tilde{I}_{WSPT}(j)+\lambda_3\tilde{I}_{ATC}(j,t), \tag{4}$$

where we have $\sum_{i=}^{3}\lambda_i = 1$ for $\lambda_i \geq 0, i=1,\ldots,3$. Here, we use normalized indices $\tilde{I}_{WSPT}(j) := I_{WSPT}(j)/\eta$ and $\tilde{I}_{ATC}(j,t) := I_{ATC}(j,t)/\eta$ where we set $\eta := \max w_j / \min p_{jl}$. Note that η is the maximum value of indices $I_{WSPT}(j)$ and $I_{ATC}(j,t)$. We are interested in determining a triplet $(\lambda_1^*,\lambda_2^*,\lambda_3^*)$ such that a compromise between throughput, the fulfillment of quality requirements, and on-time delivery performance goals is reached.

The three objectives are integrated using the desirability function approach proposed by Derringer and Suich (1980). The approach is based on the idea that each individual objective function value is transformed into a value between 0 and 1. Thus, each value of the objective function y_i to be minimized is converted into a desirability function value from $[0,1]$. If y_i meets the goal, then $d_i \equiv 1$, whereas $d_i \equiv 0$ if y_i is outside the acceptable range. We denote the maximum allowable value for the response y_i by U_i. Moreover, let G_i be the goal value for y_i. We then define the desirability function d_i that belongs to y_i by:

$$d_i := \begin{cases} 1, & \text{if } y_i < G_i \\ \left((U_i - y_i)/(U_i - G_i)\right)^{\gamma_i}, & G_i \leq y_i \leq U_i \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where $\gamma_i > 0$ is the weight of the desirability function y_i. Note that the desirability function can be formulated in a similar way when the objective function has to be maximized. In this situation, we have to

replace the maximum allowed value U_i by a minimal allowed value L_i. We obtain the combined desirability D as the geometric mean of the individual desirabilities.

Next, the desirability of the fulfillment of the quality requirements is called d_1. It is the desirability function for the throughput performance measure. The release rate is used to set the G_1 value, while the corresponding L_1 value is set by simulation experiments using the WSPT dispatching rule that is not quality-aware. Moreover, we denote the desirability of the cycle time (CT) by d_2. We choose G_2 as the average raw processing time of the products used in the wafer fab. The U_2 value is chosen by simulation experiments with the quality-aware dispatching rule since we expect large CT values in this situation. The desirability function of the on-time delivery performance measure TWT is given by d_3. We choose $G_3 = 0$. The corresponding U_3 value is set by using the TWT value for a simulation with tight due dates and the First-In-First-Out (FIFO) dispatching rule.

Overall, we use the combined desirability $D := (d_1 d_2 d_3)^{1/3}$ in the rest of the paper. D values can be derived using the dispatching rule given by the index $I(j,i,t,\lambda_1,\lambda_2,\lambda_3)$. Of course, the values of the function D depend on the concrete $(\lambda_1,\lambda_2,\lambda_3)$ setting.

3.2 VNS Scheme to Facilitate the Search for Appropriate Weight Combinations

We have to determine $(\lambda_1^*,\lambda_2^*,\lambda_3^*) = \text{argmax}\, D(\lambda_1,\lambda_2,\lambda_3)$. This can be achieved using a grid search approach. Therefore, we apply the grid $G := \{0.05(r,s,20\text{-}(r+s)) \mid r,s \in \text{IN}, 0 \le r+s \le 20\}$ in the simulation experiments. Note that we have $|G| = (21 \cdot 22)/2 = 231$ grid points in this situation. Since the grid search approach leads to large simulation burden, we are interested in expediting the search for appropriate weight combinations based on a metaheuristic. VNS is a local search-based metaheuristic that is based on the idea to enrich a simple local-search approach to avoid that the algorithm keeps getting stuck in local optima. This is achieved by restarting the local search procedure from a randomly chosen neighbor of the incumbent solution. The restarting step is called shaking. It is performed using neighborhood structures of increasing size. The basic VNS approach can be summarized according to Hansen and Mladenovic (2001) as follows:

Basic VNS

1. **Initialization:** Select neighborhood structures $N_k, k = 1,...,k_{max}$, find an initial solution x, and choose a stopping criterion. Set $k := 1$.
2. Repeat until the stopping condition is met:
   a) **Shaking:** Choose randomly $x' \in N_k(x)$.
   b) **Local Search:** Apply some local search method with x' as initial solution, and denote by x'' the obtained local optimum.
   c) **Acceptance decision:** If x'' outperforms the incumbent solution x, move there, i.e., set $x := x''$ and $k := 1$, otherwise set $k := k \bmod k_{max} + 1$. Go to Step 2.

Therefore, we start by describing how we determine the initial solution. The grid point $0.05(7,7,6) \in G$ is used. Note that this point is close to the center of gravity of the triangle that is spanned by the points $(1,0,0)$, $(,0,1,0)$, and $(0,0,1)$. The following classes of neighborhood structures are applied:

- Move(Δ,k): Randomly select two components $\lambda_i, \lambda_j, i \neq j$. If $\lambda_i \ge \Delta$ then set $\lambda_i := \lambda_i - \Delta$ and $\lambda_j := \lambda_j + \Delta$. Repeat this step k times.

- Swap(k): Randomly select two entries $\lambda_i, \lambda_j, i \neq j$ and swap the corresponding values. Repeat this step k times.

Note that the vector entries obtained by Move(Δ, k) can be different from the grid points due to the choice of the Δ value. The sequence of neighborhood structures is summarized in Table 1.

Table 1: Sequence of neighborhood structures.

Number of the Neighborhood Structure (k)	Structure
k=1,..,5	Move(Δ, k)
k=6,7	Swap$(k-5)$

The local search approach is based on moves that consider two adjacent entries of the weight vector starting from the left to the right. We set $\lambda_i := \lambda_i - \widetilde{\Delta}$ and $\lambda_j := \lambda_j + \widetilde{\Delta}$ if $\lambda_i \geq \widetilde{\Delta}$ for two selected components $\lambda_i, \lambda_j, i \neq j$. In a first pass, we always ensure $i < j$, whereas $i > j$ is ensured in the second pass. Overall, six moves have to be evaluated. The complete neighborhood of x', the vector obtained by the shaking step, is considered for a small $\widetilde{\Delta}$ value in a best-fit manner.

The overall simulation-based optimization scheme is shown in Figure 1. The dispatch module is responsible for parameterizing the dispatching rule, collecting statistics, and determining the desirability function value for a given weight vector. The VNS approach selects the next weight vector based on the desirability value from the dispatch module. The simulation model is responsible for computing lot completion times and for determining whether a lot is scrapped or not.

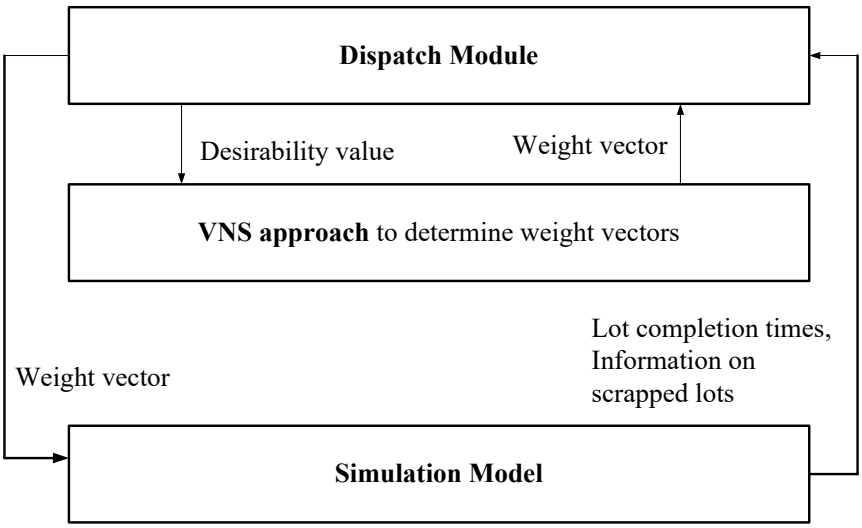

Figure 1: Overall simulation-based optimization scheme.

4 SIMULATION ENVIRONMENT

4.1 Simulation Model

We use the MIMAC I data set from Simulation Data Sets (2017) to construct a simulation model that consists of a single machine group that include five machines. The bottleneck of the MIMAC I data set, the stepper work center, is taken. The two products from the MIMAC I data set are applied to derive the routes. It is assumed that all lots include 48 wafers. Exponentially distributed machine breakdowns are used in the simulation model. Only the process steps on the steppers are modeled in detail, whereas the

process steps on the non-stepper machines are modeled as delays similar to Hung and Leachmen (1999) and Ehm *et al.* (2011). Flow factor values obtained from the full simulation model are considered to compute the lengths of the individual delays, i.e., we multiply the processing time of the corresponding process step with the flow factor value to get the mean of the delay. The delay itself is gamma-distributed. The reduced simulation model is preliminary validated against the full simulation model, i.e., we compare mean cycle time and throughput. However, it is not within the scope of the paper to get an accurate reduced simulation model out of the detailed simulation model.

4.2 Implementation Issues

Because we have to perform a large number of simulation runs for a fairly simple simulation model of a machine group we are interested in a fast simulation tool. Therefore, a simple simulation engine is coded in the C++ programming language. The simulator follows the event-scheduling approach (cf. Law 2014). The main features of the tool are the following ones:

- it allows for modeling the arrival process of the lots
- it allows for modeling the stochastic delays
- it allows for applying different dispatching rules
- it allows for modeling of machine breakdowns
- it allows for gathering various statistics.

The VNS approach is again coded based on the C++ programming language. The simulation tool is integrated with the grid search approach and the VNS approach. The simulation experiments are carried out on a computer with 2.2 GHz Intel Core i7-3632QM CPU and 8 GB RAM.

5 COMPUTATIONAL EXPERIMENTS

5.1 Design of Experiments

We expect that the due date setting has an impact on the performance of the proposed simulation-based optimization scheme. Therefore, we set the due date of lot j according to:

$$d_j := r_j + FF\, z_j \sum_{k=1}^{n_j} p_{jk}\,, \tag{6}$$

where z_j is the realization of a random variable $Z \sim U(2/3, 4/3)$. Here, we denote by $U(a,b)$ a continuous uniform distribution over the interval (a,b). The flow factor FF is determined for lot releases that lead to an average machine group utilization of 90%. Therefore, a normally distributed weekly amount of lots is generated for a product mix of 1:1. The related coefficient of variation (CoV) of the amount of lots is 0.1. The weights of the lots are considered as realizations of a random variable that is distributed according to $U(0,1)$.

Moreover, the gamma-distributed delays are generated as follows. We prescribe the CoV value. Due to $CoV = 1/\sqrt{\alpha}$, we are able to determine the first parameter α of a gamma distribution. The second parameter β is determined by $\beta = E/\alpha$, whereas the expected value E is determined by multiplying the processing time with the corresponding flow factor value.

Next, we describe how we generate the c_{ij} values. The parameters of a multi-dimensional normal distribution are determined based on real-world data from a wafer fab to model machine-specific eligibility values. Then eligibility values are randomly generated for each lot and machine. If lot j with $c_{ij} < 0.6$ is selected for processing on machine i then this lot is scrapped before it is completed with a probability of 0.1.

In the simulation experiments, 25 independent simulation replications are considered to obtain statistically meaningful results. Average values are considered for the performance measure values to determine the desirability values. A simulation horizon of 14 months is considered. This includes a warm-up period of two months.

The simulation experiments are organized in such a way that we first use the grid search approach to determine appropriate $(\lambda_1, \lambda_2, \lambda_3)$ triplets. In a next step, we apply the blended dispatching rule and the individual rules to determine throughput, TWT, and CT values. In a final experiment, we show that the simulation burden can be reduced by using VNS-based simulation optimization without compromising solution quality.

5.2 Parameter Setting

We use $\gamma_i \equiv 1, i = 1,2,3$ in all simulation experiments, i.e., we do place more emphasis on being close to the goal value and do not decrease importance on proximity to the goal value. We use the setting $\Delta = 0.01$ for the Move neighborhood structures, whereas the setting $\tilde{\Delta} = 0.01$ is used in the local search scheme. The VNS scheme is stopped if after three consecutive shaking steps no improvement is reached. Moreover, a maximum computing time of 60 minutes is allowed. The look-ahead parameter κ is chosen as follows. We carry out simulation experiments where the global ATC dispatching rule is used for a grid $\kappa = 0.5k, k = 1,\ldots,10$. The κ value that leads to the smallest TWT value is used in the experiments with the blended dispatching rule. We use CoV=0.1, i.e. $\alpha = 100$, for determining the β value in the gamma-distributed delays. The parameter settings used in the simulation experiments are summarized in Table 2.

Table 2: Parameter settings used in the simulation experiments.

Parameter	Value
weight γ_i of desirability function y_i	1.00
move length Δ	0.01
move length $\tilde{\Delta}$	0.01
CoV for gamma-distributed delays	0.1 ($\alpha = 100$)
look-ahead parameter κ	grid $\kappa = 0.5k, k = 1,\ldots,10$

5.3 Simulation results

We show the simulation results obtained by the different dispatching rules relative to the results obtained by the FIFO dispatching rule in Table 3. The throughput column is abbreviated by TH. Best performance measure values for the different dispatching rules and products are always marked in bold. Here, we abbreviate the first product by P1 and the second one by P2.

Table 3: Simulation results for the different dispatching rules relative to FIFO.

Dispatching rule	TH		CT		TWT	
	P1	P2	P1	P2	P1	P2
FIFO	1.000	1.000	1.000	1.000	1.000	1.000
ATC	0.978	1.004	0.998	0.987	0.761	**0.661**
WSPT	0.987	0.967	0.979	**0.960**	**0.752**	0.673
quality-aware	1.018	**1.095**	**0.975**	0.968	0.909	0.903
blended	**1.059**	1.089	0.995	0.975	0.778	0.679

We see that the global ATC rule leads to the smallest TWT values among the different dispatching rules followed by the WSPT rule. The WSPT rule provides the smallest average CT values, while the quality-

aware dispatching leads to the largest TH values. However, the blended dispatching rule combines the advantage of the different rules since it leads to a fairly large TH value, small CT values, and a high on-time delivery performance. The overall simulation time to determine the weight vector for the blended dispatching rule using the grid approach is 48 minutes.

Moreover, we perform the VNS approach to determine the weight vector for the blended dispatching rule. We obtain the results shown in Table 4 relative to the performance measure values obtained for the grid search approach. 62 VNS iterations are enough to compute the entries of the weight vector. This leads to an average simulation time of 12 minutes. Note that the number of grid points increases to a large extent when more then three criteria are combined. In this situation, the VNS-based simulation optimization scheme is even more important.

Table 4: Simulation results for the VNS-based approach relative to the results of the grid search approach.

Dispatching rule	**TH**		**CT**		**TWT**	
	P1	**P2**	**P1**	**P2**	**P1**	**P2**
Blended (grid search)	**1.000**	1.000	1.000	1.000	1.000	1.000
Blended (VNS)	0.966	**1.002**	**0.983**	**0.988**	**0.970**	**0.969**

We see that the performance measure values obtained by the fast VNS-based approach are close to the full grid-based approach, i.e., the difference between the blended dispatching rule obtained by a time-consuming grid search and the VNS-based approach is pretty small. Again, best performance measure values are marked bold in Table 4. The resulting weight vectors for the grid search and the blended approach are $(0.70,0.15,0.15)$ and $(0.65,0.19,0.16)$, respectively. The fairly small CT and TWT reduction of the VNS-based approach can be explained by the fact that the grid search approach uses a coarser grid.

6 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this paper, we developed a simulation-based optimization approach to design dispatching rules for a bottleneck machine group in a wafer fab that take into account both productivity- and quality-related objectives. The different objectives are integrated using the desirability function approach. On the priority index level, weights are used to sum up the different indices in order to build a blended priority index. A reduced model of a full-fab simulation model was considered to reduce the simulation burden. A VNS approach is proposed to find appropriate weight vectors within a short amount of computing time. We demonstrated that the proposed blended dispatching rule performs well with respect to the three criteria. It clearly outperforms the dispatching rules that support only a single criterion. Moreover, even for three different criteria we obtained a significant reduction of the simulation time when the simulation-based optimization approach using VNS is considered.

There are several directions for future research. First of all, a more rigorous computational assessment of the proposed method seems to be desirable. Moreover, it seems to be necessary to extend this approach towards the situation where more than one machine group is involved. While it is likely that the approach itself will work, more effort is necessary to setup the reduced simulation model in this situation. As a second future research direction it seems interesting to investigate whether the genetic programming approach proposed by Pickardt et al. (2010) can be applied to construct quality-aware dispatching rules or not. While this approach seems to be at a first glance fairly different, it is similar on a conceptual level. Moreover, it is interesting to see whether the proposed global dispatching approach can be outperformed by a more local approach where MIP approaches are used to determine the next job to be processed on a machine. Finally, we are interested in conducting simulation experiments to determine situations where changing the weight vector is a reasonable strategy.

ACKNOWLEDGMENTS

We would like to thank Chen-Fu Chien for pointing out the problem addressed in the present paper and for helpful discussions. Moreover, the authors gratefully acknowledge the support of Ying-Jen Chen who provided some real-world equipment health data sets.

REFERENCES

Chien, C.-F., Y.-J. Chen, and C.-Y. Hsu 2015. "A Novel Approach to Hedge and Compensate the Critical Dimension Variation of the Developed-and-etched Circuit Patterns for Yield Enhancement in Semiconductor Manufacturing." *Computers & Operations Research* 53:309-318.

Dabbas, R. M., and J. W. Fowler. 2003. "A New Scheduling Approach Using Combined Dispatching Criteria in Wafer Fabs." *IEEE Transactions on Semiconductor Manufacturing* 16(3):501-510.

Derringer, G., and R. Suich. 1980. "Simultaneous Optimization of Several Response Variables." *Journal of Quality Technology* 12(4): 214-219.

Doleschal, D., G. Weigert, and A. Klemmt. 2015. "Yield integrated Scheduling Using Machine Condition Parameter." In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 2953-2963, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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

Hansen, P., and N. Mladenovic. 2001. "Variable Neighborhood Search: Principles and Applications." *European Journal of Operational Research* 130:449–467.

Hildebrandt, T., D. Goswami, and M. Freitag, M. 2014. "Large-scale Simulation-based Optimization of Semiconductor Dispatching Rules." In *Proceedings of the 2014 Winter Simulation Conference,* edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 2580-2590, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hung, Y., and R. Leachman. 1999. Reduced Simulation Models of Wafer Fabrication Facilities. *International Journal of Production Research* 37(12):2685–2701.

Law, A. M. 2014. *Simulation Modeling and Analysis*. 5th ed., New York: McGraw-Hill.

Li, L., Z. J. Sun, M. Zhou, and F. Qiao. 2013. "Adaptive Dispatching Rule for Semiconductor Wafer Fabrication Facility." *IEEE Transactions on Automation Science and Engineering* 10:354–364.

Li, L., and Z. Min. 2016. "An Efficient Adaptive Dispatching Method for Semiconductor Wafer Fabrication Facility." *International Journal of Advanced Manufacturing Technology* 84:315–325.

Simulation Data Sets. 2017. "MIMAC Datasets." Accessed April 15, 2017. http://p2schedgen.fernuni-hagen.de-/indexphp?id=242.

Mönch, L., J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, and O. Rose. 2011. "A Survey of Problems, Solution Techniques, and Future Challenges in Scheduling Semiconductor Manufacturing Operations." *Journal of Scheduling* 14:583-599.

Mönch, L., J. W. Fowler, and S. J. Mason. 2013. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*. New York: Springer.

Mönch, L., and J. Zimmermann. 2004. "Improving the Performance of Dispatching Rules in Semiconductor Manufacturing by Iterative Simulation. " In *Proceedings of the 2004 Winter Simulation Conference*, edited by R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 1881-1987. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Obeid, S. Dauzère-Pérès, and C. Yugma. 2012. "Scheduling on Parallel Machines with Time Constraints and Equipment Health Factors." In *Proceedings of the 2012 IEEE Conference on Automation Science and Engineering (CASE 2012)*, 401-406.

Pickardt, C., J. Branke, T. Hildebrandt, J. Heger, and B. Scholz-Reiter. 2010. "Generating Dispatching Rules for Semiconductor Manufacturing to Minimize Weighted Tardiness." In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 2504–2515. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sarin, S. C., A. Varadarajan, and L. Wang. 2011. "A Survey of Dispatching Rules for Operational Control in Wafer Fabrication." *Production Planning & Control* 22(1): 4-24.

Vepsalainen, A. P. J., and P. E. Morton. 1988. "Improving Local Priority Rules with Global Lead-time Estimates: A Simulation Study." *Journal of Manufacturing and Operations Management* 1:102-118.

Yugma, C., J. Blue, and S. Dauzère-Pérès. 2015. "Integration of Scheduling and Advanced Process Control in Semiconductor Manufacturing: Review and Outlook." *Journal of Scheduling* 18(2):195-205.

Zhang, H., J. Jiang, and C. Guo. 2009. "Simulation-based Optimization of Dispatching Rules for Semiconductor Wafer Fabrication System Scheduling by the Response Surface Methodology." *International Journal of Advanced Manufacturing Technology* 41:110–121.

AUTHOR BIOGRAPHIES

LORENZ REINHARDT holds Bachelor and M.S. degrees in Information Systems from the University of Hagen, Germany. He works as a software engineer at höltl Retail Solutions GmbH. His research interests are in discrete-event simulation, production planning and control, and information systems for production. His e-mail address is Lorenz.Reinhardt@googlemail.com.

LARS MÖNCH is full professor of Computer Science at the Department of Mathematics and Computer Science, University of Hagen where he heads the Chair of Enterprise-wide Software Systems. He holds M.S. and Ph.D. degrees in Mathematics from the University of Göttingen, Germany. After his Ph.D., he obtained a habilitation degree in Information Systems from Technical University of Ilmenau, Germany. His research and teaching interests are in information systems for production and logistics, simulation, scheduling, and production planning. He can be reached by email at Lars.Moench@fernuni-hagen.de.