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

# DEVELOPMENT AND INTRODUCTION OF A COMBINED DISPATCHING POLICY AT A HIGH-MIX LOW-VOLUME ASIC FACILITY

Mike Gißrau

X-FAB Dresden GmbH & Co.KG
Grenzstrasse 28
D-01109 Dresden, GERMANY

Oliver Rose

Universität der Bundeswehr München
Department of Computer Science
D-85577 Neubiberg, GERMANY

## ABSTRACT

The fabrication of semiconductor devices, even in the area of customer oriented business, is one of the most complex production tasks in the world. A typical wafer production process consists of several hundred steps with numerous resources including equipment and operating staff. A reasonable assignment of each resource at each time for a certain number of wafers is vital for an efficient production process. Several requirements defined by the customers and facility management must be taken into consideration with the objective to find the best trade-off between the different needs.

In this paper we describe the development of a combined dispatching policy allowing the company to set up multiple objectives and generating the best compromise between different requirements. The rule is applied at a typical ASIC facility with a low-volume high-mix characteristic.

## 1 INTRODUCTION

A typical application specific semiconductor facility, also called Foundry, has a very complex production process. Several hundred product types and variants with several hundred process steps have to be processed on dozens of equipment. The whole production task is characterized by several re-entrant flows within the whole facility. The main tasks of a wafer fabrication process consists of the four steps patterning, layering, doping and heat treatment, in various sequences (see Figure 1). At each step, various effects on the production process are known, like equipment failures, sequence dependent setups or batching operations. These elements cause a huge variability for factory performance measures and can lead to unstable behavior.

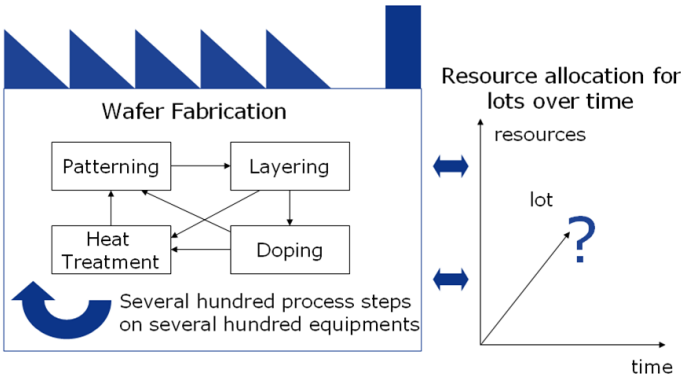


Figure 1: The wafer fabrication process.

The conditions mentioned above necessitate an intelligent approach for controlling the whole factory process. The main question to be answered is the resource allocation of the equipment and the operating staff to the production tasks as well as the order in which the different tasks have to be done. Bad decisions

can cause instabilities and a degradation of the overall facility performance. To avoid these effects a general policy is needed providing a reasonable decision at each time for each step. In the literature, a wide range of different scheduling and dispatching approaches can be found. Scheduling as described in Pinedo (2002) is quite hard to implement and use in a high-mix low-volume facility, whereas dispatching is quite common. Different techniques beginning with the definition of simple policies like in Rose (2002), Rose (2001), Rose (2003) are known. However there can also be found more complex ones (like in Dabbas, Chen, Fowler, and Shunk (2001)) taking different criteria into account, with focus on mass production. The influence of these rules is often not obvious and depends on the field of application (e.g. see Mittler and Schoemig (1999)). The complex approaches are often designed for specific areas or deal with a mass production environment. In our case there is a need for a simple approach offering large flexibility as well as good dispatching decisions at each time.

In this paper, we describe a combined dispatching approach under the usage of our simulation model described in Gißrau and Rose (2011) offering the possibility to define different objectives which are taken into consideration (see Section 2). The approach is defined in Section 2.2. Different simulation analyses are done to provide sufficient performance measures. The results are shown in Section 2.3. Based on the approach, a dispatching system is implemented offering a large flexibility in the facility environment (see Section 3).

## 2 COMBINED DISPATCHING POLICY

In this section, the combined approach is described, including the problem description and a simulation analysis.

### 2.1 The Problem Description

In the semiconductor foundry business, several requirements from the management and the customers have to be fulfilled. These requirements often affect each other. From the customer point of view, the on-time delivery of the ordered wafers and the quality are the major points. From the viewpoint of the management, each semiconductor facility has to fulfill several performance measures; besides the on-time delivery also excellent overall production parameters like cycle time (per mask layer) and a high utilization are also important. Thus for a new dispatching policy, these considerations should be taken into account.

### 2.2 The Approach

As mentioned in the previous section, a wide range of simple dispatching policies are available often optimizing just one performance parameter of interest. In our case, we use the following dispatching policies combined to a more complex one:

- **FIFO:** The First In First Out rule is the classical starting point for dispatching rule analysis. The rule offers a very low variability at all performance measure of interest and can be described as very fair. The rule takes the entry position of the lot in the equipment queue into account. The lot which is in the queue the longest time is taken next. The normed priority PR of the FIFO rule for lot $L_i$ can be calculated as
$$P_{FIFO} = 1 - \frac{i}{n} \tag{1}$$
where i is the current position of the Lot in the queue and n is the current count of all entities in the queue.
- **SPT:** The Shortest Processing Time First rule forces lots with the shortest processing time to be processed next. The objective of this rule is to maximize the throughput. This rule can cause stability problems (e.g. see (Rose 2001)) in case of variation of processing times on a highly utilized equipment. In this case, only the lots with low processing times are processed, the longer

processing time consuming lots are processed less often. Thus queue length may reach unwanted regions. The normalized priority of the SPT rule for lot $L_i$ can be calculated as

$$P_{SPT} = 1 - \frac{t_i - t_{min}}{t_{max} - t_{min}} \tag{2}$$

where $t_i$ is the processing time of $L_i$.

- **CR:** The Critical Ratio rule is widely used in factory environments and takes the due date of the lot as well as the remaining processing time of the current stage into account. It attempts to optimize two performance measures, the throughput as well as the on-time delivery. In some cases, even in highly utilized facilities, this rule tends to be unstable in case of less appropriate due date targets. The priority of the CR rule for lot $L_i$ can be calculated as follows:

$$P_{CR} = \begin{cases} N\left(\frac{1+d_{due}-d_{now}}{1+t_{RPT}}\right), & \text{if } d_i > T_{now} \\ N\left(\frac{d_i - T_{now}}{1+t_{RPT}}\right), & \text{otherwise} \end{cases} \tag{3}$$

where $N(x)$ is the normalization function of the priority values, $d_i$ is the due date of the lot, $T_{now}$ is the current date and $t_{RPT}$ is the remaining processing time of the lot.

- **EDD:** The Earliest Due Date rule is often used in semiconductor environments in an attempt to optimize the on-time delivery from a global point of view. The lot with the closest global due date at a stage is processed next. A disadvantage is the global usage of the due date does not take local considerations into account. Thus at some company profiles, the rule is not a sufficient choice. The normed priority PR of the EDD rule for lot $L_i$ can be calculated as

$$P_{EDD} = 1 - \frac{d_i - d_{min}}{d_{max} - d_{min}} \tag{4}$$

where $d_i$ is the due date of $L_i$.

- **ODD:** The Operation Due Date rule is a variant of the EDD rule defining local due dates per each lot and stage. The local due date $d_{i,s}$ at stage $s$ can be calculated as

$$d_{i,s} = \sum_{d=1}^{s} t_{p,s} XF_t = \sum_{d=1}^{s} \left(\frac{d_{due} - d_{start}}{t_{PT}}\right) \tag{5}$$

where $t_{p,s}$ is the theoretical raw processing time of stage $s$, $d_{due}$ the global due date of $L_i$, $d_{start}$ the start date of $L_i$ and $t_{PT} = \sum_{s=1}^{N} t_{p,s}$ the theoretical raw processing time of the whole process of $L_i$ with $N$ stages. The normalized priority PR of the ODD rule for lot $L_i$ can be calculated as follows:

$$P_{ODD} = 1 - \frac{d_{i,s} - d_{min}}{d_{max} - d_{min}} \tag{6}$$

- **LB:** The Line Balance rule is applied to avoid starvation of tools and tries to balance the work in process fluctuation at each stage or equipment. Figure 2 illustrates the main aim of this rule from a practical point of view. A total equality of the WIP at each stage is not possible at real factory environments due to different influences from batching operations and production variations. The objective is a more balanced work in process of the whole wafer fab which also avoids starvation of tools.

  In our case, we use two different approaches, the first one makes use of defined work in process levels per each tool group or tool. This approach is used in our simulation analysis result. Of course the determination of the goal levels is quite difficult, thus as a second solution, we can also try to balance the work load at each stage and cluster. The workload can be described as the whole time the equipment or stage needs to process all lots which are currently available.

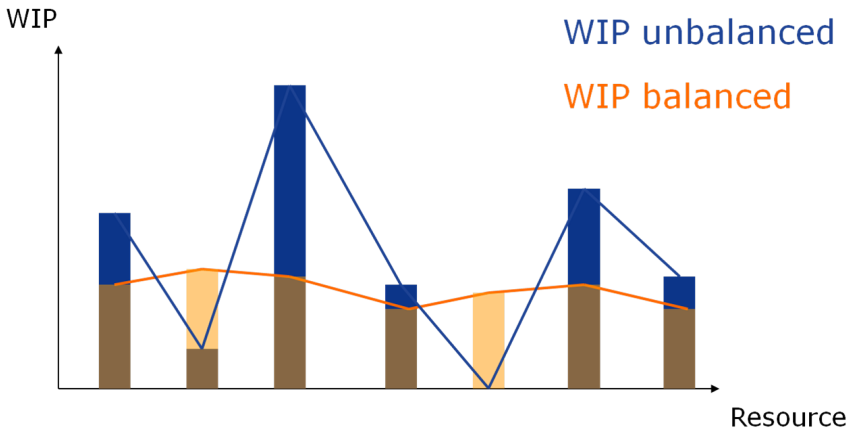


Figure 2: Line balance.

The rules mentioned above fulfill several requirements defined by the management and the customer. The throughput is optimized (with SPT), also the due dates can be taken into account (with EDD, ODD). Each of these rules is also simple to use and has a low computation effort. Now we combine these rules in a linear way with a given weight $w_k$ per rule $k$:

$$P_{Lot} = \sum_{k=1}^{K} w_k P_{Lot,k} \tag{7}$$

The weights are determined by a factory model using a genetic optimization algorithm. The whole approach is outlined in Figure 3.

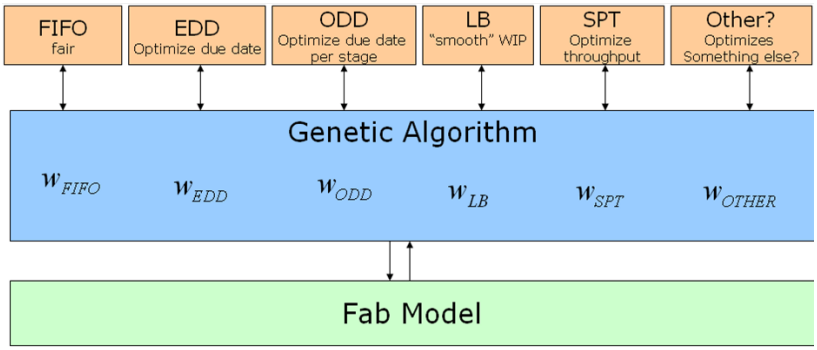


Figure 3: Overview dispatching approach.

In our research, we tested several algorithms to determine the rule weights and finally choose an algorithm based on the genetic evolution principle. Several advantages of the genetic algorithm can be shown in our approach,

- the robust nature of the algorithm,
- the simple operator usage, and
- no need for complex additional information about the system transformation function.

Our approach is fitted to the combined dispatching approach for finding the optimal weight combination $W_o$ under the given restriction $\sum w_i = 1$ and $0 \leq w_i \leq 1$ as well as a given set of dispatching policies $D = \{d_1, ..., d_n\}$. The typical encryption of the population members by a binary representation is replaced

by the vector

$$W_X = \begin{pmatrix} w_{X,1} \\ \dots \\ w_{X,n} \end{pmatrix} \tag{8}$$

where $w_i$ is the decimal value of the weight referenced to the dispatching rule $d_i$. According the genetic principle, at each iteration two members of the current population are combined to a new solution of the search space. This operation (also called crossover) is defined by the following equation:

$$F_{CROSS}(W_A, W_B) = W_A * W_B = \begin{pmatrix} (w_{A,1} - w_{B,1})z + w_{B,1} \\ \dots \\ (w_{A,n} - w_{B,n})z + w_{B,n} \end{pmatrix} \tag{9}$$

The variable $z$ is a random number with a $U(0,1)$ distribution. Figure 4 illustrates an outline of an example with $N = 3$ dispatching rules. At first, the edge points of the search space are evaluated by the system. In the case of the example, three simulation runs are necessary. The second step allocates ordered points over the whole search space to get sufficient elements for a first start population. Now the genetic algorithm calculates new population members. The green points illustrate the currently best population members, the red points worse elements, the gray points are currently under evaluation. Each population member itself is rated by the defined objective function $O(f)$. It contains the facility performance measures of interest like the on-time delivery and calculates a numerical result of each simulation run, which represents the member fitness. Each statistical parameter $p$ of the simulation with its elements like the median or the deviation can be added to the objective function in a weighted order:

$$O(f) = w_1 * p_1 + \dots + w_z * p_z \tag{10}$$

where $\sum_{z=1}^{Z} w_z = 1$. The genetic algorithm attempts to maximize this function result.

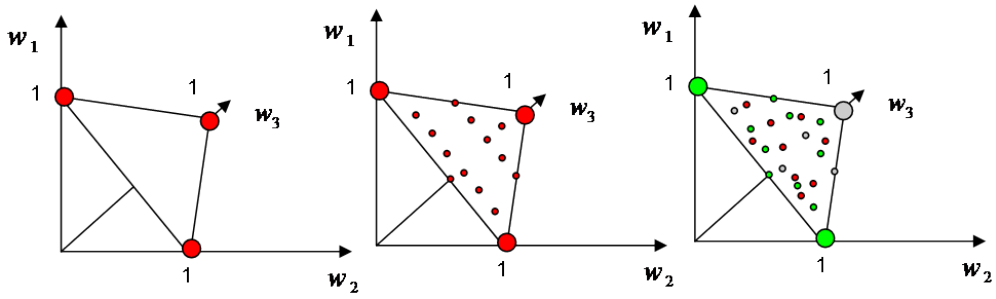


Figure 4: Example optimization run for $N = 3$ dispatching rules.

In the next section we simulate this approach for a detailed factory model (see (Gißrau and Rose 2011)).

### 2.3 Analysis by Simulation

With the detailed facility model, different scenarios were simulated and analyzed. The detailed facility model consists of about

- 300 different equipments,
- 100 different products and product variants, and
- several model capabilities like setup, batching and operator interaction.

It bases on real facility data of a low-volume high-mix ASIC facility.

The scenario presented here has been run for 6 months with historical factory data of this time period including the lot starts, the tool downs and the operator schedules. This allows us to make a comparison with the real historical behavior of the facility. We uses a JAVA based simulation environment. The number of replications is 50 in order to the low random influence in this case by usage of historical data. There is no warm up length because the model is initialized by historical state data of the start date. The current dispatching policy is named as FIFO reference, based on a extended FIFO rule. In addition we use the dispatching rules LB, SPT, ODD, EDD and classic FIFO for the weight combination. Two different weight combinations were simulated, an equal weight solution setting all $w_i = 0.2$ and an optimized version. For the optimized solution, the cycle time per mask, the average WIP as well as the on-time delivery was taken into account as objectives. Each of the objectives has the same weight at this case. Figures 5, 6, and 7 illustrate the general behavior of the different rule sets.

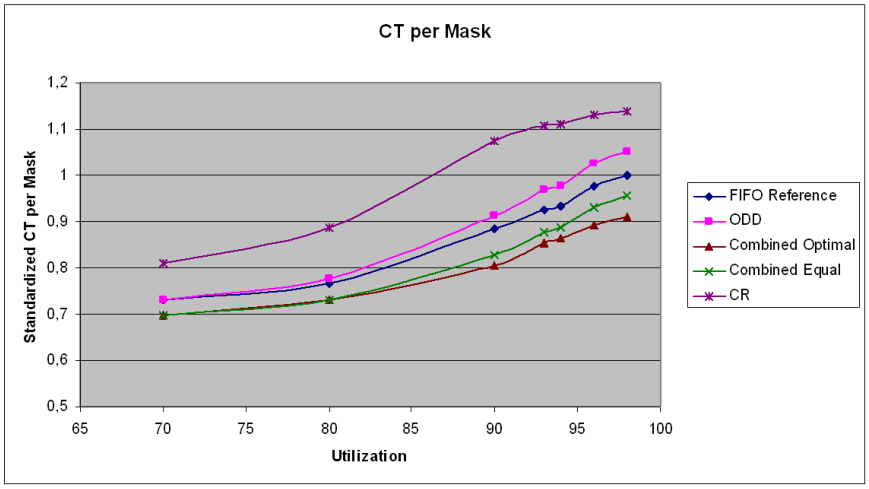


Figure 5: Cycle time per mask layer mean.

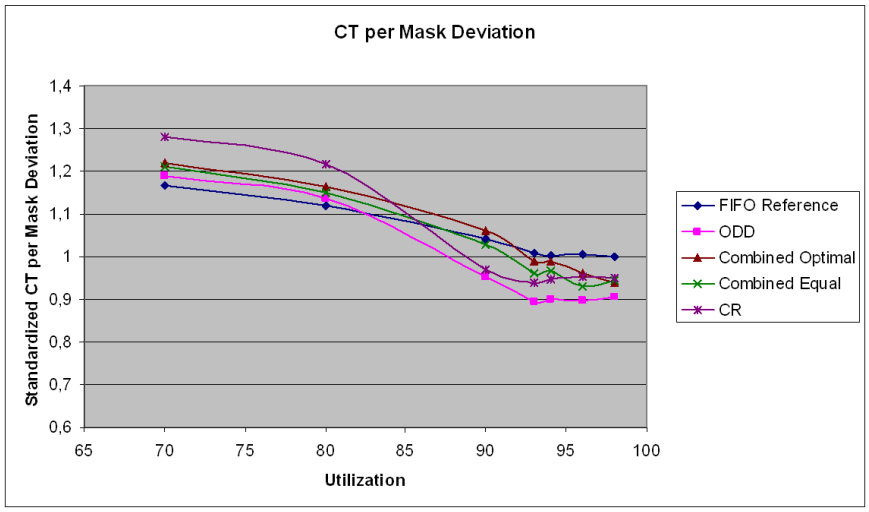


Figure 6: Cycle time per mask layer deviation.

For the cycle time per mask, we find that the higher facility utilization the larger the decrease in the cycle time per mask mean and deviation are. At maximum utilization of about 98% the average decrease

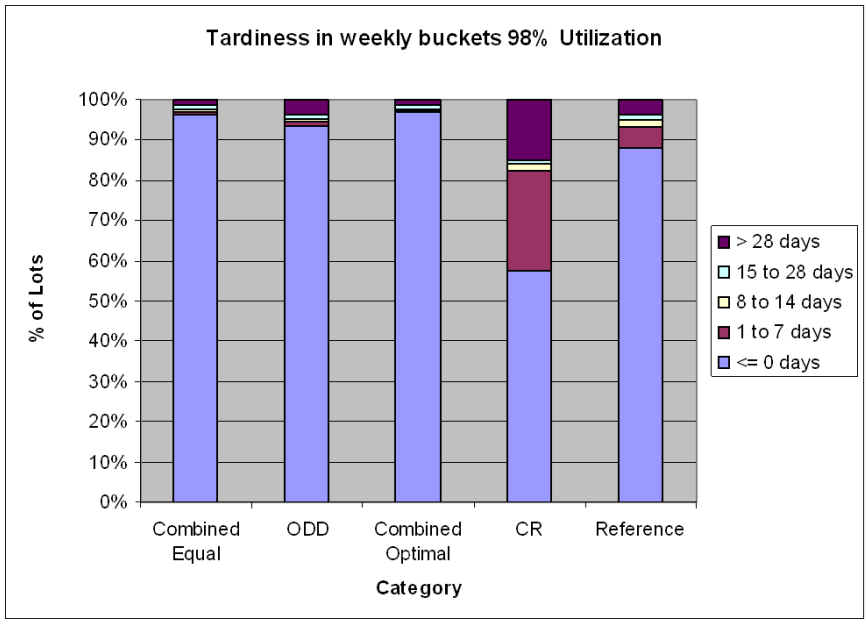


Figure 7: Average tardiness of the lots in weekly buckets.

of the mean value is about 5% for the equal weight case, and in the optimized case, the decrease is about 8%. The deviation, which is also an important performance measure, can be reduced by about 5%. All improvements are compared to the reference rule used. In our case the FIFO rule is applie. The rule contains some local optimizations like setup control and batching optimization. The tardiness of the lots, which is very important for the customer, can be improved by about 5%. In general the simulation analysis shows an improvement of about 6 to 8% for various factory performance measures in comparison to the historical reference.

### 2.4 Pros & Cons

The detailed simulation analysis of the proposed approach with real facility models offers an average improvement of 5% to 10% of several performance measures of interest, including the cycle time per mask layer and the average WIP. Even the main parameter of interest at the foundry business, the tardiness of the lots can be successfully reduced by about 5%. With the usage of an automated model generation procedure, the optimization of the dispatching weights can be used in real facility environments to ensure reasonable weight combinations under defined objectives. This procedure needs a large amount of sufficient data from the facility warehouse to generate reasonable models and to perform the different dispatching criteria, which is difficult in inhomogeneous data landscapes.

## 3 IMPLEMENTATION & INTRODUCTION

In this section, we give a short overview about the implementation of the approach to the facility environment.

### 3.1 The Requirements

Implementation and introduction of a new system in a running facility is a difficult task. The inhomogeneous IT infrastructure and the different expectations of the operating personal are barriers to be come over. For our system, we use an independent structure. To this aim the relatively new standard of web services is used, which offers a high flexibility for usage and performance aspects. Our implementation of the whole controller system is based on the JAVA programming language. Different considerations have also to be

taken into account. In our case the IT environment offers a good and widely used programming background for this language. In addition, proven concepts of application servers and clusters are established and used, which also includes web service applications. Well known best practices and documentation techniques are also available. The user interfaces are also based on common WEB 2.0 techniques without the need of local client installations.

### 3.2 System Structure

The system structure is described in Figure 8. The model generation and optimization is done in a periodical way. At dramatic facility state changes (e.g. longer equipment failures) or at certain time intervals, the optimization can be run. The dispatching part has real time characteristics. The priorities are calculated when a lot moves to the next step asynchronously, thus the list generation is a very short task for each equipment. For data access, a defined data structure is used, which is needed by the controller and the

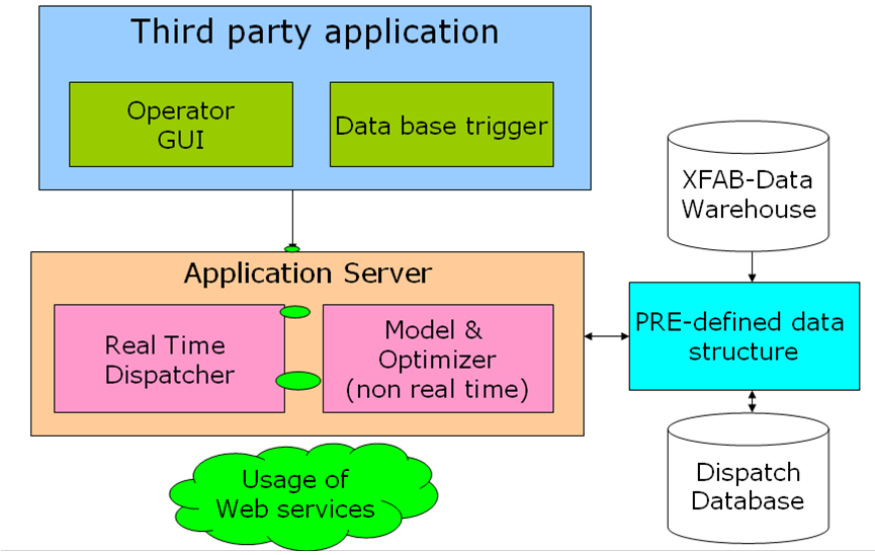


Figure 8: System overview.

model part. The structure includes current lot states, the product portfolio, the equipment states, as well as recipe data. Also operator staff data is included, containing their qualification and work schedules. The system itself is accessed via web service calls, which can be done from different systems, like non-Java based applications, but also from data base triggers if needed. The result is a very flexible structure, which can be used by very different other software components throughout the whole facility.

### 3.3 System Introduction

The introduction of the system is currently in process and will be finished by end of autumn 2012. The introduction is planned at several stages. The first tests are successfully applied at a virtual test system, which contains cloned data from the real MES and fab data storage. This allows testing without influencing the real system of the running facility. In addition different load tests are performed. These tests show sufficient performance results.

Figure 9 illustrates an example for the dispatch list generation performance test. In this case, different calling periods based on uniform random distributions with the mean values of 2s, 5s, and 10s are illustrated. In this test run the list generation needs about 2 seconds until the presentation to the user. Because of the lower performance of the test system, this time will be improved at productive system environment. At some cases due to the random method calls, simultaneous (at least 5 to 6) calls slow down the performance and

produces some outliers to the maximal time. The most time consuming actions during the list generation are the interactions with the data warehouse getting actual information like the current lot states. Improvements at this area are rarely possible due to the system restrictions not determined by the dispatching application.

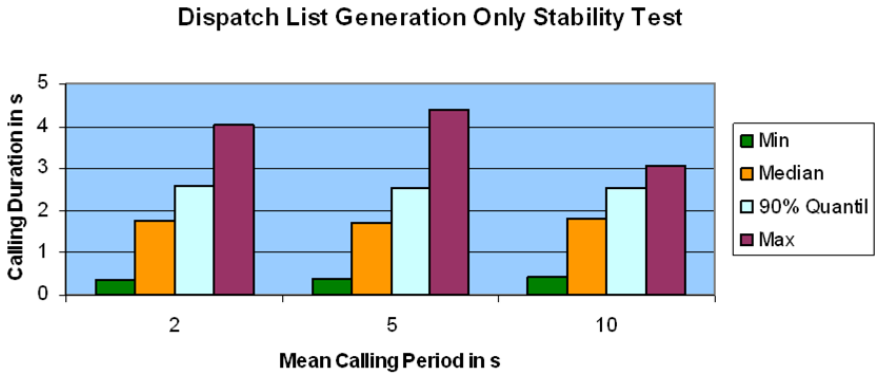


Figure 9: Stability test result for dispatch list generation.

After the successful finish of the evaluation on the virtual test system, the next steps include a productive test. The test at the productive level includes a manual comparison of the dispatch list generated by the current system and the dispatch list generated by the new dispatching controller system. This is done with an extra evaluation web site which allows to show both lists and to make comments. Also a simple rating can be done. The operating staffs have the task to evaluate this system during several weeks before the system affects the reality. After the evaluation phase a complete review of the productive test is planned in order to find weak points of the system. Finally the productive usage is planned at the end of autumn 2012.

## 4 CONCLUSIONS

In this paper, we consider the development of a combined dispatching policy for a typical high-mix low-volume ASIC facility. The combined approach shows an average improvement of the common factory performance parameters like cycle time per mask, the work in process or the on-time delivery of about 5% to 8% to the currently available reference dispatching policy based on an extended FIFO rule. We also described the system implementation based on JAVA web service technologies. This approach allows a very flexible usage and introduction of the system at several third-party applications. Also the introduction test scenario of the system is described roughly, which is one main point for the application of the system to the real production process.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Matthias Haenisch and Michael Zadlo at X-FAB Dresden for their valuable and fruitful discussions and ideas. In addition, we would like to thank the whole IT department of the X-FAB Dresden for their assistance given in the implementation phase of the project.

## REFERENCES

Dabbas, R. M., H.-N. Chen, J. W. Fowler, and D. Shunk. 2001. “A Combined Dispatching Criteria Approach to Scheduling Semiconductor Manufacturing Systems”. *Computers and Industrial Engineering* 39:307–324.

Gißrau, M., and O. Rose. 2011, December. “A Detailed Model for a High-Mix Low-Volume ASIC Fab”. In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach,

K. P. White, and M. Fu, 1953–1963. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Mittler, M., and A. K. Schoemig. 1999, December. "Comparision of Dispatching Rules for Semiconductor Manufacturing Using Large Facility Models". In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. Evans, 709–713. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Pinedo, M. 2002. *Scheduling Theory, Algorithms, and Systems*. 2nd ed. Prentice-Hall, Inc.

Rose, O. 2001, December. "The Shortest Processing Time First Dispatch Rule and some Variants in Semiconductor Manufacturing". In *Proceedings of the 2001 Winter Simulation Conference*, edited by B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 1220–1224. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Rose, O. 2002, December. "Some Issues of the Critical Ratio Dispatch Rule in Semiconductor Manufacturing". In *Proceedings of the 2002 Winter Simulation Conference*, edited by E. Yücesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes, 1401–1405. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Rose, O. 2003, December. "Accelerating Products under Due Date Oriented Dispatching Rules in Semiconductor Manufacturing". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 1346 – 1350. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

## AUTHOR BIOGRAPHIES

**MIKE GIßRAU** is a PhD student at the Universität der Bundeswehr München. He is member of the X-FAB Dresden Semiconductor Facility responsible for the innovation of the new dispatching system. He received his M.S. degree in Computational Engineering from Dresden University of Technology. His research interests include different dispatching concepts, optimizations and their realization in factory environment of complex production facilities. His email address is Mike.Gissrau@xfab.com.

**OLIVER ROSE** holds the Chair for Modeling and Simulation at the Department of Computer Science of the Universität der Bundeswehr München, Germany. He received an M.S. degree in applied mathematics and a Ph.D. degree in computer science from Würzburg University, Germany. His research focuses on the operational modeling, analysis and material flow control of complex manufacturing facilities, in particular, semiconductor factories. He is a member of IEEE, INFORMS Simulation Society, ASIM, and GI, and General Chair of WSC 2012. His email address is Oliver.Rose@unibw.de.