PRACTICAL ASSESSMENT 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 practical assessment of a combined dispatching policy presented in Gißrau and Rose (2012). Besides the facility performance influence, also the human factor is taken into consideration. This includes dispatch compliance parameter and staff surveys.

1 INTRODUCTION

A typical application specific semiconductor facility, also called Foundry, has a very complex production process. A wide range of different process steps at a characteristic re-entrant material flow makes wafer fabrication very complex. Different known and unknown influences likes equipment failures or operator staff breaks cause a huge variability for factory performance measures and can lead to unstable behavior. The wide diversification of different requirements at the foundry business like a stable on-time delivery and short cycle times necessitate an intelligent approach for controlling the whole factory process.

In the literature, a wide range of different scheduling and dispatching approaches can be found, often with focus on mass production or academic examples. 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), and 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)).

We introduced a combined approach (see Gißrau and Rose (2012)) where the different needs could be combined to a flexible and adaptable solution for practical use. The practical assessment of the four month test run at a high-mix low-volume ASIC facility is presented in this paper.

2 SHORT REVIEW

In this section, we introduce the combined policy as a short review of the paper (Gißrau and Rose 2012).

2.1 The Combined Dispatching Policy

In the semiconductor foundry business, several requirements from the management and the customers have to be fulfilled. These requirements often affect each other. Major points are for example the on-time delivery of the lots or a appropriate cycle time (per mask layer). When optimizing one parameter, a negative influence of the other parameter is possible. The combined dispatching policy consists of a variable set of different single-objective dispatching policies. The goal is to find the best trade-off between the different requirements for the fab.

For our research, we apply the following range of simple dispatching policies which often just optimizing one performance parameter of interest:

- **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 lot which is in the queue the longest time is taken next. The normed priority P 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. 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 attempt to optimize two performance measure, 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. 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. 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.

The rule set is combined in a linear way to find the best compromise between the different needs by assigning a weight w_k to each rule k and calculating the resulting lot priority P_{Lot}:

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

The determination of the weights is realized by a detailed facility simulation model. The simulation model is feed by data from the facility data warehouse including different data sources like MES or ERP system. A genetic algorithm calculates the weight for each dispatching rule by a given objective function. This objective function is defined by the management and logistic department and changes over time, depending on the requirements of the different customer. Figure 1 illustrates the whole approach. Several simulation experiments are done to justify the impact of the different weight combinations as well as the relationships between the rules for high-mix low-volume ASIC fab characteristics. For a deeper look into detail we refer to Gißrau and Rose (2012) and Gißrau and Rose (2011).
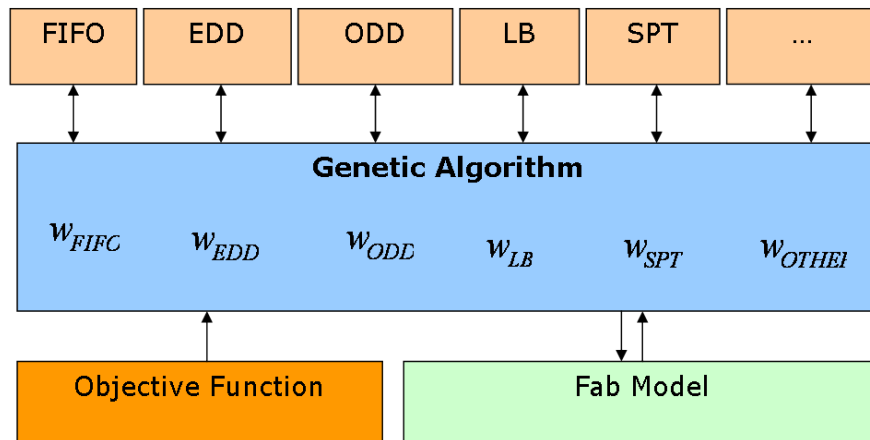

Figure 1: Overview dispatching approach.

2.2 System Implementation

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 personnel are barriers to be come over. Even the IT infrastructure offers a wide range of different data sources required for the simulation model and the real-time dispatching. Figure 2 illustrates the general environment.
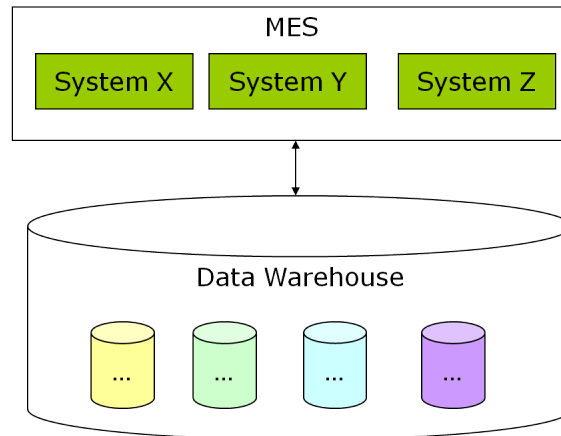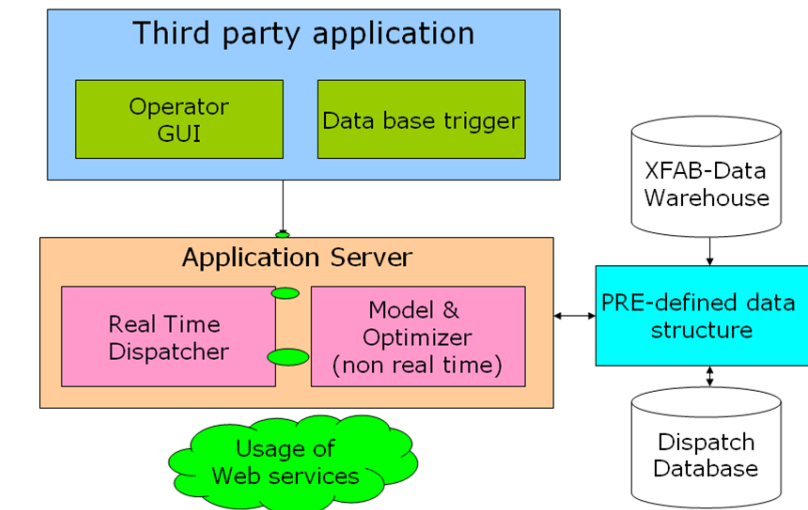

Figure 2: System environment.

Figure 3: System overview.

The fast change of different software systems at our facility force us to use an independent application structure. To this aim the relatively new standard of web services is used, which offers a high flexibility for usage and performance aspects. Figure 3 gives an impression of the whole system implementation.

The system is based on the JAVA programming language. Standardized interfaces allows multiple access from other third party applications like the operator interface. The system is divided into

- **Real Time Dispatcher:** The real time dispatcher is responsible for dispatch list generation for the different equipments and the lot priority update at lot movements. Due to the real time characteristics, time consumption at the different method calls should be very low.

- **Simulation and Optimization Core:** The simulation and optimization core is responsible for calculating the new weight combinations according the genetic algorithm. For this a detailed facility model is generated from the data warehouse.

Because of the inhomogeneous data landscape, a standardized data interface is defined. The interface is filled by data from different data sources, like the MES system or files. The data actuality ranges from on-line data in case of lot positions and states to frequently updated data in a certain interval (1h to 24h).

3 PRACTICAL ASSESSMENT

In this section, we illustrate the practical results of the application of the combined approach to a high-mix low-volume ASIC facility. During the test period of four month (End of August 2012 to end of December 2012), the dispatch list at each equipment are generated by the combined approach. At our evaluation, the most important parameter presented here are like the WIP (Work in Process) or the CM (Cycle Time per Mask Layer - the cycle time divided by the number of mask layers for a better comparableness between the different technologies) are collected with their statistical values like the mean, the median (50%) or the 90% quartile.

3.1 Facility Performance Influence

During the evaluation period, the dispatching system has a positive influence on the facility performance behavior. Figure 4 illustrates the absolute performance development over the past month with the mean values of the operator count, the WIP, the wafer out and the 50% and 90% quartile of the CM. Due to a low facility load, improvements are not clearly detectable for the average cycle time per mask and the 90% quartile.
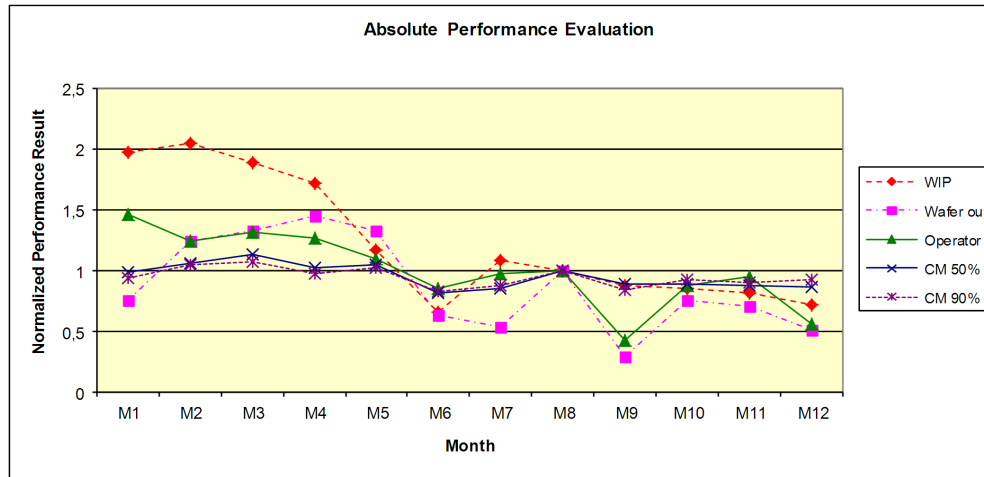

Figure 4: Absolute performance evaluation.

A look into the cycle time per mask layer distribution illustrate the positive influence of the new approach. Figure 5 shows the absolute S-Curve for the cycle time per mask layer without any normalization. The interesting quarters of the year 2012 are Q3 and Q4. In Q1 and Q2, the technology mix and the facility load changed, therefore only Q3 and Q4 could be compared, where the boundary conditions were the same.

The positive effect takes place at Q4, due to the average processing duration of two to three month for one lot. At the lower border, the increases of the curves in Q3 (first effects of dispatch control) and Q4 are much sharper than in Q1 and Q2. That indicates a better management of faster and slower lots. Before the dispatch algorithm starts, lots are sometimes delivered to early whereas lots with a strength due date

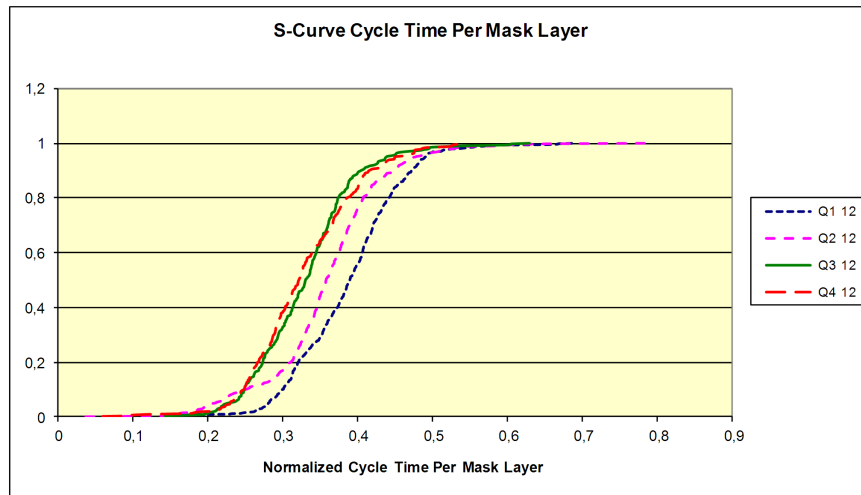**S-Curve Cycle Time Per Mask Layer**

Figure 5: S-Curve cycle time per mask for all technologies - no normalization.

were delivered to late. A further effect could be detected around the median of the curve. The absolute value could be successfully reduced by 5%. The worse values at the upper end of the curve are caused by production problems rather than by the dispatch control system in this case.

Figure 6 illustrates S-Curve of the cycle time per mask layer with normalization of the absolute values. Often, different technologies have different cycle time per mask layer distributions due to processing reasons. By elimination of these effect, the positive effect becomes more obvious. Besides a sharper increase, also the distribution values of Q4 are lower than of Q3.
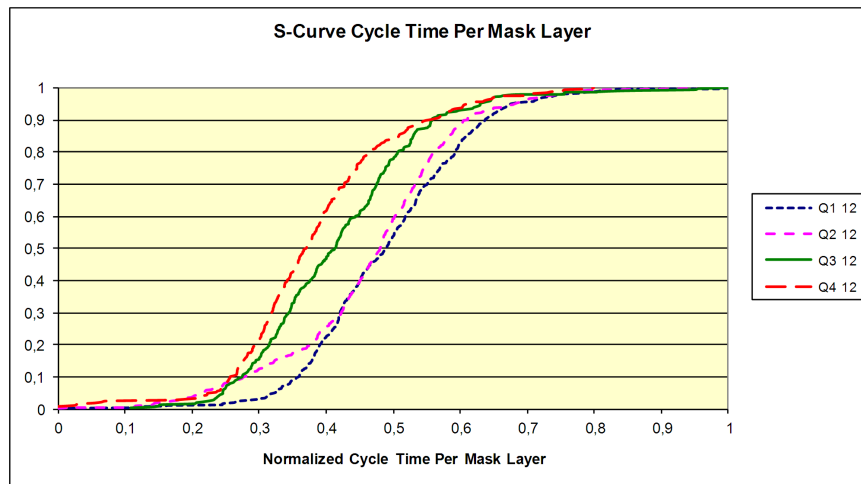
S-Curve Cycle Time Per Mask Layer

Figure 6: S-Curve cycle time per mask for all technologies - with normalization.

Besides the whole technology mix, a detailed analysis of the different cycle time per mask layer distributions of each technology is done. For the main technology, the improvements range between 5% to 10% at various levels. A shaper increase of the S-Curve at Q4 indicates a lower variation for the resulting values than in Q3.

In general the introduction of the combined approach shows a positive influence of the distributions of various facility performance measures. Even the reduction of the deviation and the reduction of extreme values (to early - to late lots) shows the expected behavior.

3.2 System Performance Evaluation

The system performance of the system is vital for the acceptance of the system by the operating personnel. To that end several performance parameters are analyzed during the evaluation period of the system. Even the time until the dispatch list is presented to the user is one of the main performance measures. Figure 7 illustrates the statistics for this case.
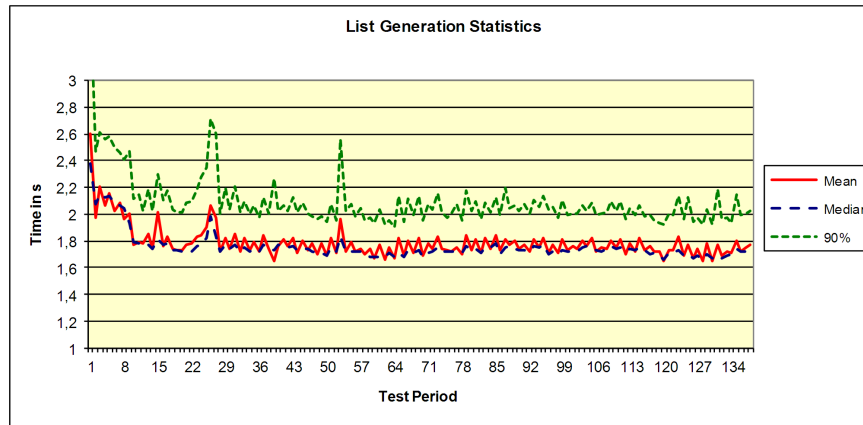

Figure 7: List generation statistics.

The median value is about 1.7 seconds, whereas the 90% value is around 2 seconds. The 90% value represents the main statistical output for our needs. Values above 3 seconds are not accepted by the operating staff. The most amount of time is caused by detecting the right lots for the equipment rather than by the dispatch algorithm itself. Various optimizations like the usage of materialized views reduce the time amount required.

The calculation of the lot priorities at a lot movement is done in a asynchronous way. Therefore the time requirements are not as strict as in the list generation case. After some improvements at the start of the test period, the 90% value ranges about 1.4 seconds, which is a acceptable value (see Figure 8).
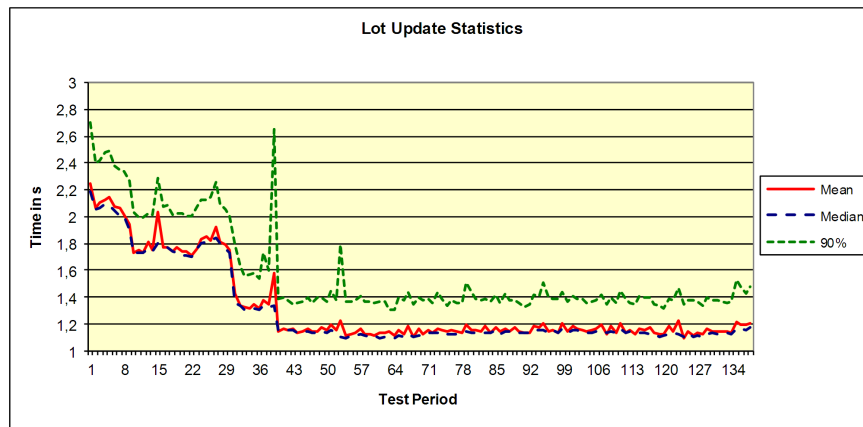

Figure 8: Lot priority recalculation statistics.

In general, the system performance is restricted by the surrounding IT infrastructure rather than by the dispatch algorithm itself. Time consumptions could be successfully reduced to a acceptable region.

3.3 Dispatch Compliance and Staff Survey

Besides the system performance, the dispatch list compliance is an important indicator for the acceptance of the system by the operating staff and the fulfillment of all requirements by the dispatching system. Thus, several statistical analyses were carried out to obtain an overview about the compliance development at the facility. The parameter i describes the list position of each lot in a dispatch list with N lots. We define different compliance parameter:

- **Compliance Score:** The compliance score P_C is a declarative value for determination of the overall dispatch compliance:

$$P_C = \begin{cases} 1, & \text{if } i = 1 \\ \frac{N-i}{N}, & \text{if } 1 < i \leq N. \end{cases} \tag{8}$$

  The value ranges between 0 and 1, whereas the target value is 1.
- **Absolute Dispatch Compliance:** The absolute dispatch compliance P_A describe the absolute fulfillment of the dispatch list:

$$P_A = \begin{cases} 1, & \text{if } i = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

  The target value is 1. In case of batch tools, the first batch fulfill the requirements for $P_A = 1$.
- **Average Sort Index:** The average sort index P_S describes the average lot position taken from the list:

$$P_S = i. \tag{10}$$

- **Correlation:** For evaluation purposes, the correlation P_{corr} between the waiting time T_{Lot} of a lot and it's priority P_{Lot} is an interesting measure how the priority of a lot affects it's waiting time:

$$P_{corr} = Corr(T_{Lot}, P_{Lot}). \tag{11}$$

  In general, we expect a negative correlation. The higher the lot priority, the lower the waiting time of the lot. The correlation value range between -1.0 and 1.0.

The compliance parameters are collected after each 12 hour work shift. At our case, four teams are available switching between night and day shift. Figure 9 illustrates the compliance score at the test period. Figure 10 shows the correlation value for the four shifts over the test period.

The compliance score ranges between 0.55 and 0.68 at the test period, with a positive trend to a higher dispatch compliance. The committed target for this value is $P_C = 0.75$. Several reasons for only a partial fulfillment of this value are known:

- Manual batch transport of lots between clusters: Lots are physically not available at the next operation but waiting for transport due to the manual transport system. The information of the current physical lot position is not available.
- Work organization: Several small storage buffer in front of the equipments lead to preference of FIFO processing. Operator tend to use lots in their direct environment rather than taking lots from storage places far away.
- Equipment characteristics: Cluster tools requiring special order of lots according to the recipe properties. Often cluster tools are not fully represented at the MES system. Information to several process areas in the tool are only partially available.
- Equipment state: Sometimes certain recipes or lots requires a conditioning of the tools. These conditioning action only has a defined life time. The current conditioning state is not available for the system. Therefore operator personnel may take lots which have the correct condition rather than change tool conditioning.
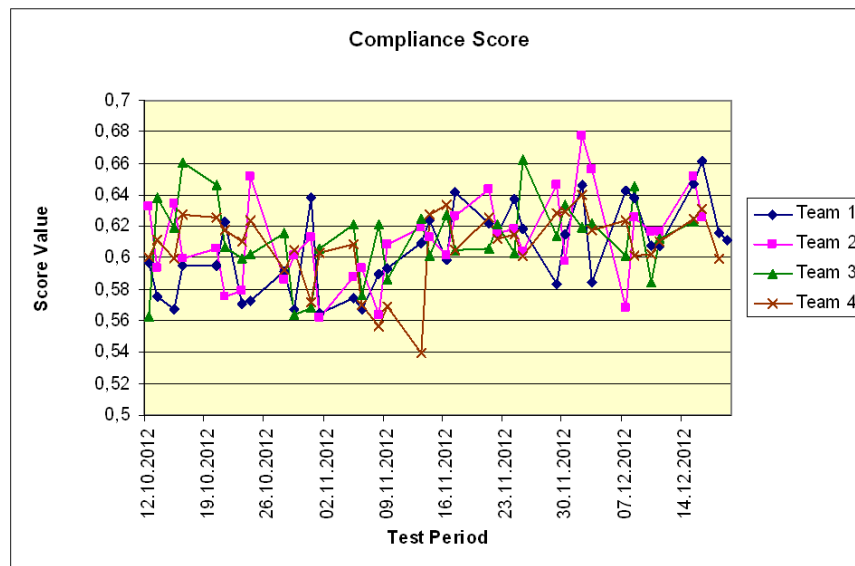
Figure 9: Compliance score per shift.

The correlation between the waiting time of the lot and it's priority shows a small negative correlation with a negative trend. We expect a higher negative correlation. Reasons for the low negative correlation could be found at low utilized equipment where only a small amount of lots are processed. Often operator personnel is not available at these tools, therefore single lots with higher priority has to wait. A more negative correlation is expected by a higher dispatch compliance.

In general, the compliance analysis offers several improvement possibilities according the work organization and the available data. Several projects are started to increase the dispatch compliance, started by operator training and data source improvements.

3.4 Staff Survey

After the end of the test period, a web based staff survey is realized to obtain an overview about the personnel opinions from leading staff. The questions are divided into four areas:

1. Evaluation of the current dispatching procedure;
2. Opinions about improvements and required changes for the current dispatching system;
3. Evaluation of the new available dispatching procedure; and
4. Explanatory notes.

The period for the survey is defined to 14 days. The survey is attended by 12 persons . These persons include the shift leader and the dispatching personnel. Operators do not participate, due to reduced working hours and the small number of operators available at each cluster. The following points show the main results of the survey:

- Current dispatch procedure: 11 persons characterize the FIFO policy and the manual priorities as not sufficient for the foundry business with high-mix low-volume characteristics. All persons range the existing policy with a general rating worse than 2 (1perfect, 6worst case), 5 worse than 3
- Required changes and needs: All persons rate the need of automated production control systems as very important. 11 persons see the need for a new production control approach
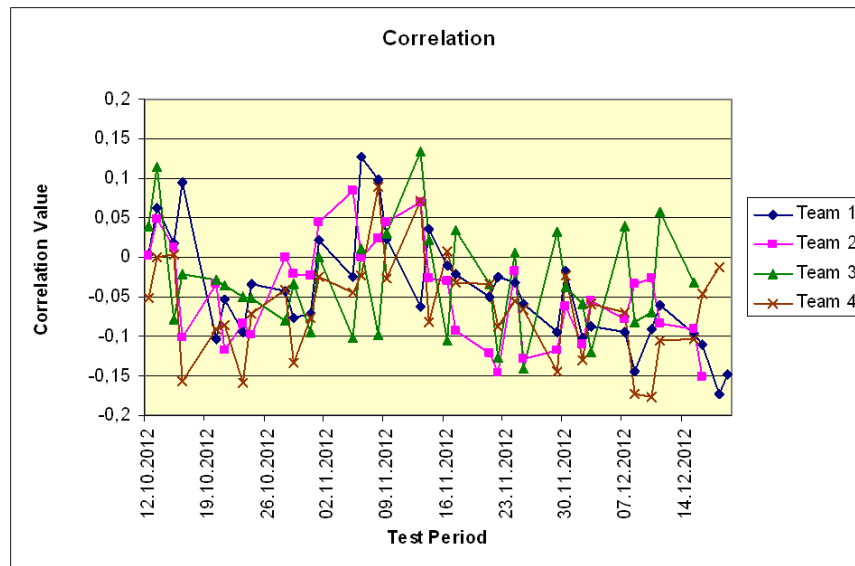
Figure 10: Correlation per shift.

- New dispatch approach: 9 persons evaluate the dispatching lists as more reasonable. 8 persons see the new approach as a useful improvement. 10 persons range the new policy with a general rating better than 3 (1...perfect, 6...worst case)

In general, all interviewed persons see a general improvement with regards to the existing solution. In addition to the results of the dispatch compliance, the survey also shows improvement potential at a few local work stations as well as the work organization. A general awareness for the importance of a reasonable production control exists. Every person acknowledges a larger improvement in regards to the existing policy definition.

4 CONCLUSIONS

In this paper, we consider the practical assessment 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 3% to 5% to former quarters during the test period. The system performance collect during the test period shows sufficient time consumptions during the method classes like dispatch list generation. The dispatch compliance analysis shows improvement potential in the area of operator training, work organization and data source management. In general the proposed dispatch approach becomes a accepted tool for a more efficient production control in our fab.

ACKNOWLEDGMENTS

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. "A Detailed Model for a High-Mix Low-Volume ASIC Fab". In *Proceedings of the 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.

Gißrau, M., and O. Rose. 2012. "Development and Introduction of a Combined Dispatching Policy at a High-Mix Low-Volume ASIC Facility". In *Proceedings of the Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 2027–2036. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Mittler, M., and A. K. Schoemig. 1999. "Comparision of Dispatching Rules for Semiconductor Manufacturing Using Large Facility Models". In *Proceedings of the Winter Simulation Conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. 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. "The Shortest Processing Time First Dispatch Rule and some Variants in Semiconductor Manufacturing". In *Proceedings of the 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. "Some Issues of the Critical Ratio Dispatch Rule in Semiconductor Manufacturing". In *Proceedings of the 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. "Accelerating Products under Due Date Oriented Dispatching Rules in Semiconductor Manufacturing". In *Proceedings of the Winter Simulation Conference*, edited by S. Chick, P. J. Snchez, 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 Wurzburg 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.