SIMULATION-BASED SYNTHESIS OF COMPOSITE DISPATCHING RULES

Martin Schickmair
Martin Graml
PROFACTOR Produktionsforschungs GmbH
Im Stadtgut A2
4407 Steyr-Gleink · Austria
E-Mail: martin.schickmair@profactor.at

Christoph Pichler
ATENSOR Engineering and Technology Systems GmbH & CoKG
Im Stadtgut A2
4407 Steyr-Gleink · Austria
E-Mail: christoph.pichler@atensor.com

Walter Laure
Infineon Technologies Austria AG
Siemensstraße 2
9500 Villach · Austria
E-Mail: walter.laure@infineon.com

KEYWORDS

Dispatching, scheduling, simulation, optimisation, lead time, model, semiconductor manufacturing, wafer test, fab, eM-Plant.

ABSTRACT

This paper describes the simulation-based synthesis of optimum dispatching rules for the wafer test area of a semiconductor production plant. The target of the optimisation task is the minimum average lead time under the constraint of maintaining on-time delivery. Dispatching is a very fast and robust technique for sequence-planning even when cycle times are not perfectly controlled. A broad range of standard dispatching rules can be found in the literature, each with its own strengths and weaknesses. Less information is available on the use of combinations of these rules. An analysis of the properties that influence the lead times and delivery dates of the lots in the facility shows that composite dispatching rules need to be used to meet the specified target.
In this paper we describe how simulation is used to explore the design space of composite dispatching rules. Based on a simulation model of the entire wafer test area, we have tested various combinations of dispatching rules under realistic operating conditions. The optimum rule combination thus found leads to an average lead time reduction of 15% while maintaining, or improving, on-time delivery.

INTRODUCTION

The design and fabrication of an integrated circuit is a complicated endeavor (Boning 1991). Modern semiconductor technology uses fabrication process sequences that consist of several hundred process steps. An integrated circuit (IC) is created from a raw semiconductor wafer by sequentially applying a specified processing sequence to the wafer. Wafers are moved through the fabrication facility (fab) in lots of 20 to 100 units. In general, more than one processing station is qualified to perform a given processing step in a product's processing sequence. Product lots are routed from one processing station to the next according to station availability. Process flows are reentrant, i. e., a lot may pass through the same group of processing stations up to more than 20 times according to the process technology used.
At the end of the processing sequence, which may take several weeks to complete, basic device functionality is verified at the parameter control measurement (PCM) step. Finally the ICs are electrically tested against product specifications on the test floor. Those circuits that do not

Proceedings 16th European Simulation Symposium
György Lipovszki, István Molnár © SCS Press, 2004
ISBN 1-56555-286-5(book) / ISBN 1-84233-106-x(CD)

pass the test are flagged and the tested wafer is shipped to the customer. Due to the large number of ICs per wafer and the complexity of the circuitry, the final testing of an entire product lot may take more than a week.
The development of new fabrication processes and new product generations continuously pushes the limits of knowledge of the physics involved. Fabrication facilities and equipment involve large investments that are amortized by fabricating a large number of different product families at a given facility. Fabrication facilities, in general, are not purpose-built for a specific product. Rather, product and process technologies are developed for existing facilities.
The facility considered in this paper comprises more than 1000 processing and testing stations and handles between 500 and 1000 different products and several hundred process flows in several process technologies simultaneously. Total output is about 30000 wafers per week. Work-in-progress inventory (WIP) is a significant cost factor, as is equipment utilisation. Demand is extremely volatile both in total volume and in the mix of the products. An effective dispatching strategy is crucial for keeping WIP low, meeting delivery schedules, and maintaining responsiveness to order fluctuations. The average run time of product lots through the facility is an excellent indicator of how well these criteria are met.

SITUATION ANALYSIS

In Figure 1 the complete process that is performed in a production line is shown. Beginning with the arriving of a job in the plant it makes its way through the manufacturing area and the backend of line to the parameter control measurement (PCM).

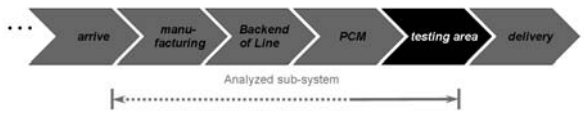

Figure 1: Principle process flow: from silicon to a finished wafer

When a lot passes the PCM it is carried to the wafer test area (WTA). The emphasis of this work is on the WTA at the end of the production cycle; this area is therefore discussed in more detail.
The WTA consists of more than 100 testers. Dependent on the product type of the specific lot it can be processed on one to about 30 different testers. This number of variability is called the "flexibility" of the lot.
Whenever a tester finishes processing the current lot a green light on top of the machine indicates to the operator (skilled worker in the WTA) that it is ready for the next lot. The operator then uses a terminal placed in the vicinity of the tester to determine which of the lots that are currently waiting can be processed on this specific machine. He is presented with a sorted list of lots and he is supposed to take the one on the top. The sorting of the lots is done by the dispatch software based on specific rules (dispatching rules).
Based on these rules the sequence of the lots through the WTA is decided. This has enormous effect on the average lead times and the adherence to delivery dates.
As we will show, a broad range of rules and combination of rules are principally applicable, each with its specific strengths and weaknesses. To identify which rules contribute most to the given objective is quite a complex and delicate problem. Thus the problem was solved using a detailed simulation model of the testing area as test-bed for assessing and optimizing different rules.

OBJECTIVES

The overall objective of our project was to identify, develop and optimise new dispatching rules for the testing area. Target of the optimisation process was to reduce the average lead times of the lots through the WTA under the constraint that the adherence to delivery dates has to be at least the same as today.
The lead time of a lot (time span from entering the WTA until leaving it) is the sum of all waiting times before processing and all physical processing times. It is obvious that the physical processing times cannot be reduced by dispatching rules, so only waiting times can be affected by. In fact dispatching rules may change lot sequences which yield in shorter waiting times.
Most time several lots are waiting for the same resource, respectively a tester. These lots differ in their properties, for example:

- their flexibility (on how many testers in the whole WTA can it theoretically be processed).
- their process time (the time it takes to perform all the necessary tests).
- their delivery date – some lots have to be delivered earlier than others.
- set-up: Does the specific machine need to be set-up, or is the lot tested before of the same type as one of the currently waiting ones?

Because of these lot-specific properties the sequence of lots does affect the average lead time and the adherence to delivery date. Therefore an ideal dispatching rule has to consider all this information in the right way. How the properties influence the objective and how the "right" balance can be found is shown in the following sections.

APPROACH AND ASSUMPTIONS

Now we know that our rule consists of the right combination of the lot-specific properties; dynamic, discrete com-

puter simulation is the right way to meet this challenge. Why?

- We have to regard that the lots in the WTA affect each other - lots are fighting for the same resource at the same time – the WTA is a very dynamic system.
- Static analysis has already been shown to be insufficient to meet our goals.
- Similarly without any reliable test-bed human intelligence and expertise turns out to be insufficient to deal with that amount of complexity.
- The product mix of the fab (which lots are in the WTA present at the same time) needs to be considered for finding the rule – exact quantification of the effects can only be done with simulation.
- The real process is not available for experiments as it operates 24 hours 7 days a week. Any application of dispatching rules that are not fully tested would bear too much risk.

As we need only ***relative*** comparisons of different dispatching rules, we can reduce the effort for data collecting and modelling by allowing some simplifications and abstractions between real world and the model. It is not the aim of the project to predict absolute production values as well funded predictions of future orders are not available anyway. It is expected that rules showing better performance with past data will also promise a performance increase for future orders. The model will be based on the following assumptions:

- The availability of the testers is not affected by dispatching rules, so we don't have to consider machine break-downs in our model.
- The same applies to transportation times of the lots from one machine to another – in the simulation model the transport times are zero.
- In reality the plant is operated by a limited number of machine operators. So if several machines need to be set-up at the same time additional waiting times may occur due to resource bottlenecks. In our perfect simulation-world we ignore these operator based bottlenecks and assume to have an unlimited amount of operators available. The relative comparison is nevertheless exact.

The following chapter describes how the model was implemented.

THE SIMULATION MODEL

The model was built using the discrete event simulator eM-Plant™ in version 7.0. EM-Plant was developed as SIMPLE++ (Simulation in Production Logistics and Engineering programmed in C++) by Tecnomatix, Stuttgart (Germany). Today eM-Plant is a standard software in the automotive industry for object-oriented graphic and integrated modeling, simulation and optimisation. Complex systems and business processes can be represented in a realistic way. The advantages of conventional concepts such as modules, language and lists are integrated in eM-Plant. EM-Plant is a single simulation system for production, logistics and engineering; these are the reasons for choosing it.

Following some details about using eM-Plant for our problem are described. Our model consists of entities, modules, tables and methods. Entities represent the lots travelling through the plant. All characteristics of a lot are assigned to the entity using custom attributes. For the intermediate buffers between the machines the standard module "storage" from eM-Plant is used; the "station" module is used to model the testers.

The main task of model implementation was to reproduce the correct flow of the lots through the wafer test area, which is controlled by the specific dispatching rules. Each lot is attributed with a product type (over 1000 different products) and the number of wafers within the lot (between 1 and 50). Lots are generated by an order source randomly based on distributions retrieved from historical production data. Each lot generated waits in a storage until a tester becomes available and the dispatcher clears the lot for processing. The wafers are tested for a determined time which usually varies according to the product type. If no further inspection is needed the lot exits the testing area and leaves for shipping.

In our model the interval for order arrival is based on an exponential distribution. The values for the product type and the number of wafers are sampled from empirical distributions based on data analysis of the real WTA. Controlled by these distributions the entities (lots) are created and their attributes are assigned. Typical attributes are product type, entering-time, number of wafers and a table of the process sequence specified for the specific product type (see window in figure 2). The entries for a process step in the sequence table are process time, the code for this process step and a table of suitable working stations (testers).

After assigning the basic attributes the entity is sent to a storage, where an entrance control calls a user defined method (methods see figure 2 on the left side). This method writes the type of the entity and some other relevant information into a custom attribute at each station where the specific lot can be processed. This attribute is called the "virtual-buffer" and is realised as a table containing all lots that can be processed on this specific station. Doing so for each station the "virtual-buffer" contains all actual lots that are waiting for processing.

Another method searches for all potential testers (see figure 2 middle) waiting for a lot and moves the lot the specific station.

Whenever a station finishes processing another method is called which chooses the next lot from its virtual buffer based on the actual dispatching rules.
The tested lot is moved to the exit method as the new one starts processing. In the exit method all relevant statistics are written to internal tables. At the end of a simulation experiment the collected data are written to Microsoft-Excel for further data processing.

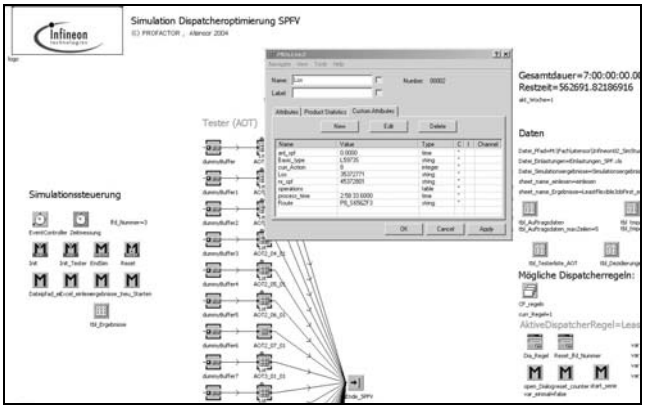

Figure 2: Screenshot: Part of the simulation-model

To execute series of simulation runs we have implemented an auto-experimenter. This module refreshes experiment-dependent variables between every simulation run (experiment) and writes the results via ActiveX in a predefined Excel table, from which several charts and experiment summaries are automatically generated. Statistics of interest are for example minimal-, maximal- and average door-to-door time, maximal- and average WIP, utilisation for each station, number of tested lots per station, maximal numbers of entries in the virtual buffer per station or calculation time per run.
Based on the simulation results we developed and optimised the new dispatching rules for choosing the next lots from the virtual buffer.

RESULTS

First we want to show the possibly not so obvious positive effects that the consideration of the processing time of a lot in the dispatching rules can have. Figure 3 shows a scenario where two lots are competing for a tester. Let us assume that lot 1 has a processing time of 1 time unit, lot 2 of 10 time units. In this simple case two sequences are possible: If we process lot 2 first an then lot 1 the sum-lead time of the both lots calculates to 21 time units. For the second possible sequence – first lot 1 and then lot 2 the sum is only 12 time units. The difference is 43% with the higher value as base! This example shows clearly that the optimised dispatching rule has to regard among other parameters the processing times in that way that the lot with the shortest test time should be processed first (Shortest-ProcessingTimeFirst – SPTF). Another and probably the most important positive effect of SPTF is that the average number of lots waiting in the WTA is reduced (see also (Rose 2001)).

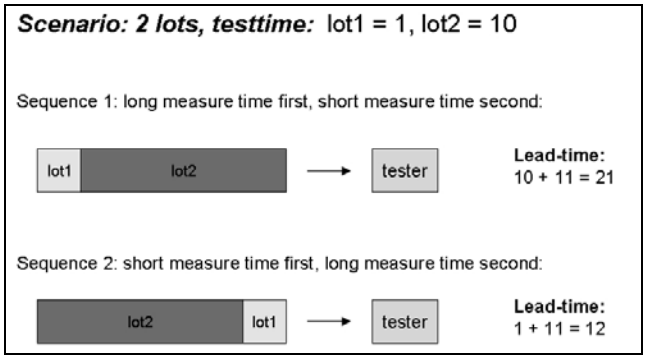

Figure 3: Effects of ShortestProcessingTimeFirst (SPTF)

But unfortunately the situation is not that easy: Without considering other lot-specific properties SPTF performs rather poor. It does not optimise the adherence to delivery dates and leads to a very unbalanced tester-utilisation that in turn results in increasing waiting times.
As already mentioned, the combination of the lot-specific properties to a single rule is the key for success. This proofed to be true after a large number of simulation experiments. Based on heuristic optimisation and far reaching discussions with operators and process experts we achieved our aim. A weighted linear combination of the lot-specific properties provided the best results.

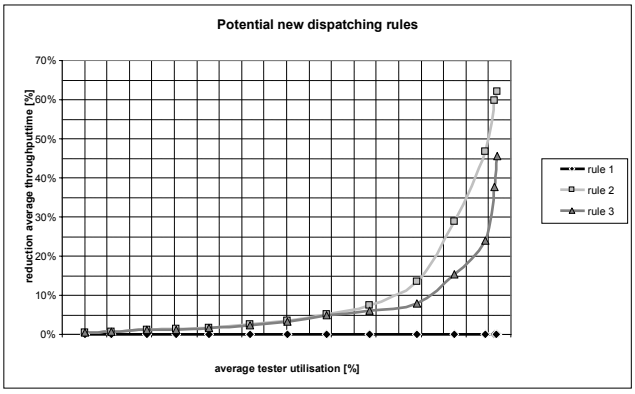

Figure 4: Lead time reduction with optimised dispatching rules

In figure 4 we show three graphs, each representing a specific dispatching rule. On Y-axis we have put the lead-time reduction in percent, on X-axis the average tester utilisation. Unfortunately but understandably we are not allowed to present the exact dispatching rules and values of utilisation. But the effects are clear: Compared to the

"old" rules currently implemented the optimised rules perform better, the higher the average tester utilisation is.
The reason for this behaviour is found in the fact, that an increasing plant utilisation leads to a greater average number of waiting lots in the intermediate buffers and therefore to an increased spectrum of possible solutions for choosing an "ideal" lot. For every rule the average lead-times grow with increasing values of utilisation, but using the optimised rules with a lower gradient.
The positive effects of the optimised dispatching rules for the real WTA can be summarized as following:

Lead time:

- at the average tester-utilisation of the real WTA a reduction of about 15% is achieved!
- the relative reduction increases heavily with higher tester utilisations (see Figure 4)

Adherence to delivery

- guaranteed at least equal to today's values
- in average better because of shorter lead times

Average number of lots in WTA (WIP)

- 15% less WIP caused by 15% shorter lead times

Investment necessary to implement

- no invest in new hardware necessary
- it is estimated that the implementation of the new dispatching rules in the WTA will need an effort of about one personnel-day

However, one interesting question is still not answered so far: How near to the theoretical optimum perform the new dispatching rules? Or the other way round, how much potential is still available for better rules?
To answer this question we have made the following estimation: We took a representative lot-arrival-sequence to the WTA with a duration of three weeks and fed it into our simulation model. During simulation execution the sequence of the lots through the WTA were logged using the old dispatching rules. Then the new dispatching rules and a heuristic optimisation algorithm based on Simulated Annealing were used to optimise the order sequence in order to reduce the average lead time to a minimum. Because of the long calculating time we allowed the heuristic algorithm to perform, it is quite sure that the found sequence is very close to the theoretical optimum. The results are shown in figure 5.
The figure shows two Gantt-Charts representing the sequence of the lots on a section of some stations. One line in the chart stands for one tester, each rectangle for a specific lot. The darker a rectangle on the chart represents the longer the waiting time of the specific lot. As expected we see that the sequence based on dispatching produces more "dark-lots" than scheduling does and it takes longer time to finish processing all orders.

The formerly developed optimised dispatching rules minimise the average lead times by 15%, the scheduling even up to 41%!

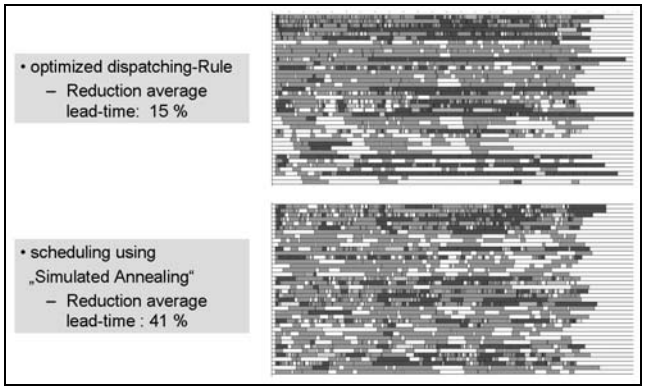

Figure 5: Dispatching vs. Scheduling (Simulated annealing)

Now we know that our sequences generated by our dispatching algorithms are far away from the theoretical optimum, but what are the reasons for that?
The answer is quite clear: Dispatching chooses the lot to be processed next under the number of lots that are actually waiting in front of the specific tester and has to decide its solution based on general rules considering only local information. Simulated Annealing does not decide on specific rules, but it tests a vast number of sequences and takes the best it finds. Furthermore the scheduling algorithm was programmed and optimised to reach the theoretical optimum in a way to consider the lots that will arrive in future too. Dispatching cannot use this information because it is impossible to predict exact arriving times of lots in the real WTA.
On the other hand in the real WTA the scheduling algorithm cannot be implemented because of the limitations found in real processes like machine break-downs, unknown or not exactly known process times and long calculation times. Anyway the result is a good estimation on how good our new dispatching rules perform.
Finally the project-team could convince the plant manger responsible for the WTA to implement the new dispatching rules in the real facility.

CONCLUSION AND FUTURE WORK

Dispatching is a very fast and robust technique for sequence-planning. A broad range of dispatching rules can be found in literature, each with its own strengths and weaknesses.
The most attractive rule for a given problem has always to be customized to the special facility and the given objectives and requirements. Our project has shown that all lot- and facility-properties that influence the lead-time and the

adherence to delivery have to be combined in the right way to reach the theoretical optimum as near as possible.
Human intelligence and expertise is insufficient for dealing with that amount of complexity. Even simple combinations of rules have surprising effects. It has been shown that computer-simulation is a very comfortable, efficient and maybe the only tool for finding the optimal solutions for a real facility. The return on investment of the presented project is excellent.
The next logical step for the project team is to optimise the dispatching rules beyond the wafer test area, for the whole fab. To find the optimum for this problem it seems to be necessary to include one more dynamic property of the facility, namely the actual number of waiting lots on the downstream stations (see figure 6).

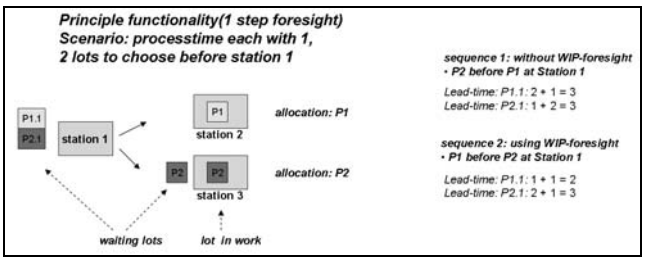

Figure 6: Lead time reduction with optimised dispatching rules

The figure above shows a simple scenario of two lots waiting before station 1. Lot "P1.1" has successor "Station 2" and "Lot 1.2" the "Station 3". If we proceed lot "P1.1" the average lead times are shorter than otherwise. This effect has to be considered by the dispatching rules.
First simulation experiments have shown that finding the optimal combination is very critical but not impossible. The project-team is confident to reach the goal again.

REFERENCES

Atherton, L.F. and R.W. Atherton. 1995. Wafer Fabrication: Factory Performance and Analysis. Kluwer.

Atherton, R.W; F.T. Turner; L.F. Atherton; and M.A. Pool. 1990. "Performance Analysis of Multi-Process Semiconductor Manufacturing Equipment." In Proceedings of theIEEE/SEMI Advanced Semiconductor Conference 1990.

Banks, J. (ed).: Handbook of Simulation – Principles, Methodology, Advances, Applications, and Practice; John Wiley and Sons, 1998; ISBN 0-471-13403-1.

Benecke, C.: „Simulation von Materialfluss- und Lagerprozessen". Zeitschrift für Logistik, Heft: 5, 1990, S. 30 - 32

Boning, D. S. 1991. Semiconductor Process Design: Representation, Tools, and Methodologies. PhD thesis, Massachusetts Institute of Technology, 1991.

Domschke W. Armin Scholl, and Stefan Voß. Produktionsplanung. Springer, 1997.

Fowler, J. and J. Robinson. 1995. Measurement and improvement of manufacturing capacities (MIMAC): Final report. Technical Report 95062861A-TR, SEMATECH, Austin, TX.

Kosturiak, J., Gregor, M.: Simulation von Produktionssystemen. Springer Verlag, 1995

McKiddie, R., Brown, S., and Neacy, E. 1994. Predicting the Impact of Short-Term Increases in Wafer Starts on a Constant Start-Rate Semiconductor Factory: Applications of SEMATECH's Future Factory Analysis Methodology. SEMATECHTechnology Transfer 94022223A-XFR..

Pichler, C. 1997. Integrated Semiconductor Technology Analysis. Österreichischer Kunst- und Kulturverlag.

Rose, O. 2001.The Shortest Processing Time First (SPTF) Dispatch Rule and Some Variants in Semiconductor Manufacturing. In Proceedings of the 2001 Winter Simulation Conference, pp. 1220-1224.

Zeichen, G., Fürst K.: Automatisierte Industrieprozesse. Springer Verlag, Wien, New York 2000

AUTHOR BIOGRAPHIES

MARTIN SCHICKMAIR, born in Wels, Austria studied electrical engineering at the Technical University Vienna and received a degree of a Dipl.-Ing. He joined the simulation based design and optimisation department at Profactor Produktionsforschungs GmbH in 1999. His special interests are the integrated simulation of technical and business-processes as well as simulation based optimisation of production logistics.

MARTIN GRAML, born in Linz, Austria studied Mechatronics at the Johannes Kepler University Linz. He is with Profactor Produktionsforschungs GmbH since 2003. His special interests are in the field of simulation based optimisation and software development.

CHRISTOPH PICHLER heads ATENSOR's Holistic Engineering department, focusing on the optimum design and operation of production facilities in the automotive and electronics sectors. Before joining ATENSOR, he was with National Semiconductor Corporation in Santa Clara, where he led the company's concurrent engineering efforts group. He holds a Ph.D. degree from Vienna University of Technology and an M.B.A. from INSEAD. His professional interests include the organisation and operation of manufacturing companies and the deployment and integration of digital methods across the enterprise.

WALTER LAURE joined Infineon Technologies as a Software Project Engineer in 1996. For four years he has been responsible for line simulation at Infineon Technologies. His career started with Alcatel as a software developer in the High Speed Network area. He has a diploma in Technical Mathematics awarded by the Technical University in Graz, Austria. His professional interests include wafer fab simulation, line logistics and developing and integrating software in semiconductor wafer fabs.