Proceedings of the 2007 INFORMS Simulation Society Research Workshop

IMPROVING THE ACCURACY OF SIMPLE SIMULATION MODELS FOR COMPLEX PRODUCTION SYSTEMS

Oliver Rose

Institute of Applied Computer Science
Dresden University of Technology
Dresden, 01062, Germany

ABSTRACT

Semiconductor wafer fabrication facilities (wafer fabs) are among the most complex production facilities. A large product variety, hundreds of processing steps per product, hundreds of machines of different types, and automated transport lead to a system complexity which is hard to understand and hard to handle. For educating planners and developing adequate material flow control mechanisms, simple models for this complex environment are required. We present several scenarios where a very simple model is used to mimic the behavior of a complex wafer fab. These studies show both the strengths and weaknesses of applying models with a high level of abstraction compared to the real world system. Recently, we found an approach to overcome some of the weaknesses

1 INTRODUCTION

The main reason for the increased use of simulation in the operational planning lies in the size, complexity, and cost of nowadays semiconductor fabrication facilities (fabs) generated by market and business pressures coupled with the hard limits of physics. Traditionally, many operational decisions in the industry were made based on prior knowledge, experience, and intuition. This is no longer appropriate. There is a need to build a meaningful model of the factory and to perform simulation studies to examine its operational problems. At the moment, there is no other analysis tool available that is capable to support meeting production goals while avoiding unnecessary investments or other costs.
Simulation is used in such areas like capacity planning, scheduling, bottleneck identification, impact of new products or process flows, layout analysis, equipment modeling, factory ramp-up modeling, and operator modeling. Typical performance measures are cycle time, throughput, inventory levels, equipment usage, and cost.

In most cases, industrial engineers in the semiconductor industry work with very detailed simulation models. There are scenarios, however, where these large models cannot be longer used due to their enormous runtimes. In particular, if the behavior of the factory over time has to be analyzed or if the fab simulation model outputs (for instance, product cycle times for a given product mix) are required as an input for higher-level enterprise performance models, e.g. supply chain models. In these scenarios, simple and fast models are required to make the analysis feasible.

The paper is organized as follows. In the next sections, we will outline some modeling approaches and results from previous studies (Rose 1998, Rose 1999a, Rose 1999b, Rose 2000). These results will show that simple models have positive effects on understanding certain fab phenomena but that for a lot of practical cases the accuracy of their predictions is not adequate. Then, we will present the improvement approaches where we are currently working on and provide some first (and promising) results. This is ongoing work and at the workshop, we will be able to present even more results.

2 PREDICTING FAB BEHAVIOR OVER TIME

In numerous studies (e.g., Wein 1988) the long-term behavior of the fabrication facilities in terms of mean cycle times, average inventory levels, etc. is determined. These studies help to find dispatch rules for achieving given requirements such as a certain probability to meet due dates. There are fab phenomena, however, that cannot be explained with such classical simulation approaches because only long-term or steady-state performance criteria are taken into consideration. One of these phenomena is the observation of huge amounts of work in progress (WIP) even weeks after a catastrophic failure of the bottleneck work center, i.e., all machines of the work center that constrains the fab capacity are down for a few days. This par-

ticular fab behavior was reported by several semiconductor fab managers.

In contrast to classical simulation studies, we need to study the evolution of the fab, for instance the WIP over time, after the catastrophic event and not the long-term behavior of the fab. To this end, we apply simulation techniques that were used to compare the performance of routing algorithms in communication networks after a node breakdown. The major disadvantage of such techniques is the fact that instead of tens of simulation runs for classical studies, hundreds of them are required for studies of the system behavior over time.

Hence, we first have to develop a simple fab model that shows the behavior of a complete fab model with respect to our problem. To carry out the study with the complete fab model is not possible due to the enormous run length of hundreds of simulation replications.

Due to the layered nature of semiconductors, the wafers visit sequences of machines several times, i.e. they proceed through the fab in cycles. Memory chips may have up to 60 layers. This cyclic visiting sequence of machines is responsible for a large part of the logistic problems of wafer fabs because lots with different due date requirements compete for the machines. If due-date based dispatching is applied, lots that are closer to their due dates are preferred at the cost of waiting time for the other lots.

To make a simulation study feasible with respect to running time, we require a fab model that shows the aforementioned behavior, but is considerably less complex in terms of the number of machines. Fig. 1 shows the proposed factory model. It consists of a bottleneck work center, a delay unit, and a control unit. The bottleneck work center determines the fab performance to a large extent (Atherton and Atherton 1995) and is therefore modeled in detail considering the number of machines, processing times, and dispatch rules. The rest of the machines are modeled as a delay unit. Each time a lot leaves the bottleneck work center it is delayed for a random amount of time before it either leaves the fab or it requests a bottleneck machine once more. The control unit decides whether the required number of layers/cycles have been finished, and directs the lots to the fab exit or back to the bottleneck work center.

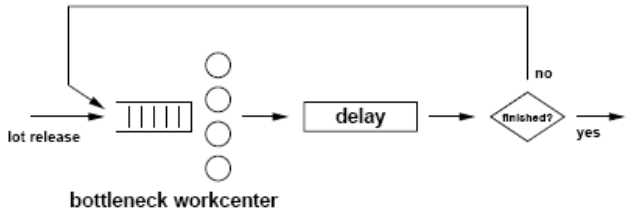

Fig 1: Simple Model

With this model we run several experiments, e.g., with different delay time distributions, with different dispatching rules at the bottleneck work center, etc. The details can be found in the original paper.

Fig. 2 shows the inventory (WIP = Work In Progress) over time after the breakdown for several dispatching rules and shifted exponential delays.

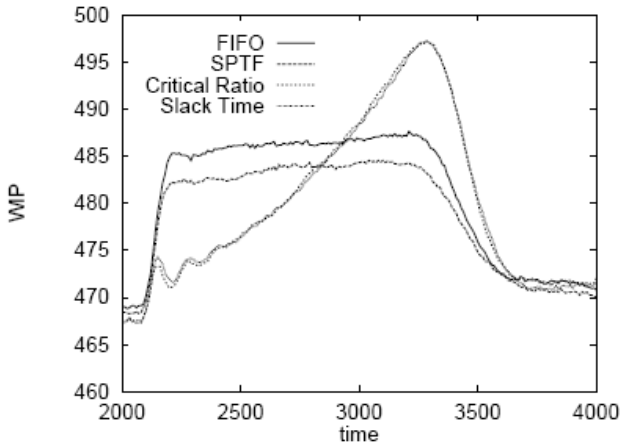

Fig. 2: Inventory Over Time After Serious Breakdown

It turns out that this behavior is same as for real wafer fabs. The slow start and the considerably larger maximum of the inventory level for due-date dependent dispatching rule can be explained by the re-entrant material flow in combination with priority changes during dispatching according to due dates.

In summary, this simple model provided an explanation for an unexpected fab behavior to the industrial engineers who introduced this problem to us.

3 PREDICTING CYCLE TIMES

Due to the success of our first study, we tried to find new application scenarios for our simple model. We identified cycle time distribution prediction as a useful one because these predictions are needed in a lot of higher level enterprise planning approaches.

We enlarged our model to include delays for the time period before entering the bottleneck for the first time and the time period after leaving the bottleneck for the last time (Fig. 3).

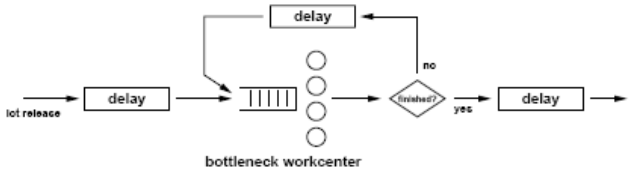

Fig 3: Enlarged Fab Model

It turned out however, that we were not able to match the distribution of the cycle times of the full detail and the simple model (Fig. 4).

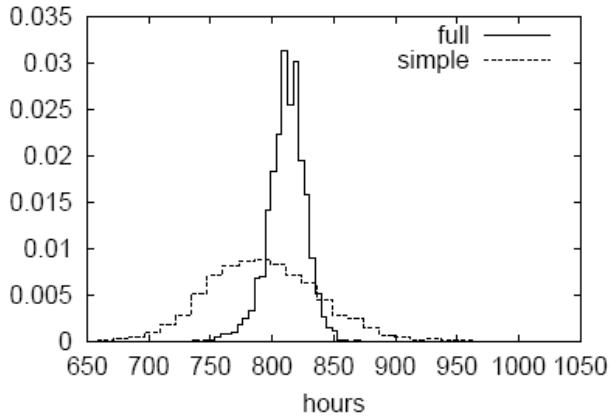

Fig 4. Product Cycle Time Distribution Deviations

The main reason for that lack in modeling capabilities lies in the fact that lots can pass each other in the delay term. In contrast to the simple model behavior this does almost never happen in the real fab or in the full detail models. Fig. 5 and 6 show this phenomenon. In Fig. 5, the per-layer cycle times of the full model are clearly separated.

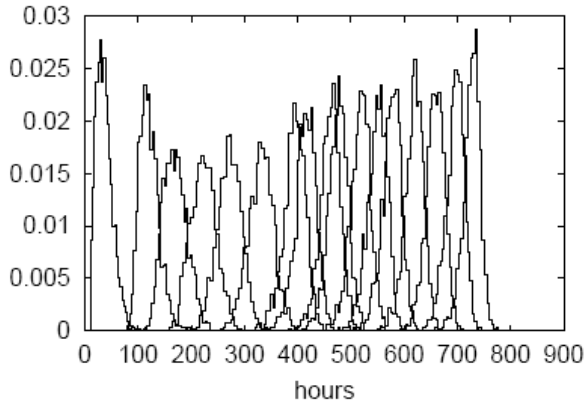

Fig. 5. Per-layer Cycle Times of the Full Model

In Fig. 6, however, the per-layer cycle time distributions become broader and broader from layer to layer.

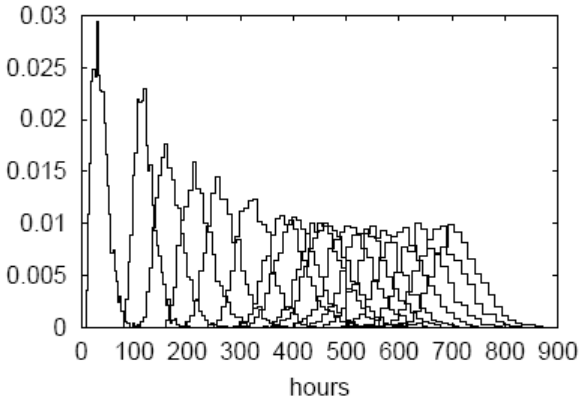

Fig 6. Per-layer Cycle Times of the Simple m Model

In addition to this weakness, the simple model can only be used for the bottleneck utilization it was calibrated for. It is not possible to use this model to generate an estimate for the characteristic curve of the fab, i.e., the cycle time over bottleneck utilization curve. In Fig. 7, we use the flow factor instead of the cycle time, where the flow factor is the cycle time divided by the sum of raw processing times.

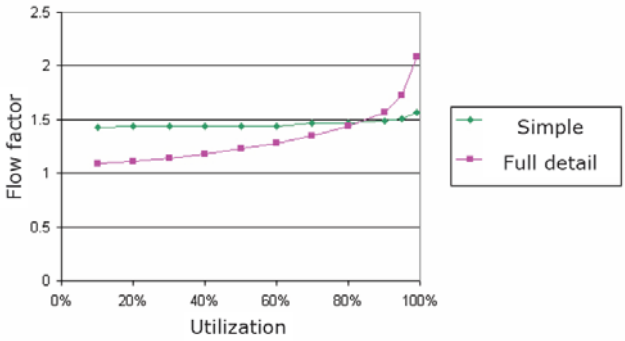

Fig. 7. Comparison of Characteristic Curves

We developed and tested a variety of simple model improvements but did not achieve considerable improvements in the accuracy of the predictions.

4 RECENT IMPROVEMENTS

In a new study, however, we considered another approach to increase the model dependency on bottleneck utilization changes. The main problem is that the fab utilization is defined by the sum of the busy periods of the bottleneck divided by the total time available for processing, i.e. this value can only be determined for a time period of reasonable length. As a consequence, we cannot determine the current utilization directly. Therefore, we decided to use an indirect way to measure the current utilization of the fab. From our perspective, the easiest way to do that is to

determine the current inventory/WIP level. In our case, it is sufficient to use the number of lots which can be found in the bottleneck work center and the center delay unit. We name this value “lots in loop”.

To make the model utilization dependent, we replace the fixed delay time distributions in the delay units by delay time distributions that are depending on the current inventory level. We perform a very long simulation run with the full detail model where we gradually increased the fab utilization from 1% to 99% in order to obtain delay time measurements for the complete utilization range. The result of this experiment is depicted in Fig.8.

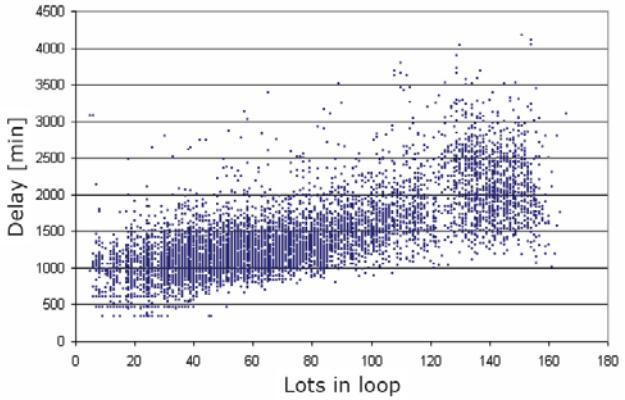

Fig 8. Inventory dependent delays

Based on these delay measurements, we determined a sequence of delay time histograms, where each histogram is related to a certain range of lots in loop. In almost all cases, the empirical delay time distributions are very close to shifted exponential distributions.

The simple model with load dependent delays has a characteristic curve that matches the curve of a full detail model much better than the original simple model (Fig. 9).

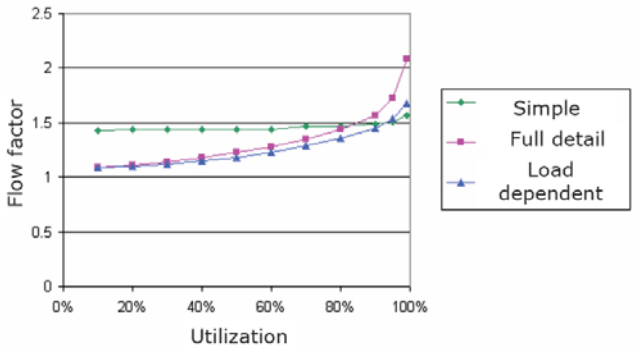

Fig 9. Improved characteristic curve

It turned out, however, that the cycle time distributions of the improved simple model are still much broader than those of the full detail model.

5 SUMMARY AND OUTLOOK

In recent experiments, we were able to find a very simple model that mimics the load dependent behavior of a full detail model. The simple model is also capable to mimic the inventory trajectory of a real factory.

At the moment, we are working on a modification of the simple model to solve the cycle distribution problem. Our current approach is to replace the center delay unit by a single server with load dependent service times. By the time of the workshop, we will be able to present first results.

ACKNOWLEDGMENTS

The author would like to thank all students who worked on the simple model projects so far and, in particular, Ralf Sprenger who is currently performing the simulation runs and the result data analysis.

REFERENCES

Atherton, L. and R. Atherton. 1995. *Wafer Fabrication: Factory Performance and Analysis*. Boston: Kluwer.

Rose, O. 1998. “WIP evolution of a semiconductor factory after a bottleneck work center breakdown.” In *Proceedings of the Winter Simulation Conference '98*, 997-1003

Rose, O. 1999a. “Estimation of the cycle time distribution of a wafer fab by a simple simulation model.” *In Proceedings of the SMOMS '99 (1999 WMC)*, 133-138.

Rose, O. 1999b. “CONLOAD - A new lot release rule for semiconductor wafer fabs.” In *Proceedings of the Winter Simulation Conference '99*, 850-855.

Rose, O. 2000. “Why do simple wafer fab models fail in certain scenarios?” In *Proceedings of the Winter Simulation Conference '00*, 1481-1490.

Wein, L.M. 1988. „Scheduling semiconductor wafer fabrication.” *IEEE Transactions on Semiconductor Manufacturing, 1(3)*:115-130.

AUTHOR BIOGRAPHY

OLIVER ROSE holds the Chair for Modeling and Simulation at the Institute of Applied Computer Science of the Dresden University of Technology, 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. Web address: <www.simulation-dresden.com>.