Proceedings of the 2007 Winter Simulation Conference
S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.

IMPROVED SIMPLE SIMULATION MODELS FOR SEMICONDUCTOR WAFER FACTORIES

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. Several years ago, we published some first approaches which were useful to explain the fab behavior after a serious bottleneck breakdown. With that simple model, however, it was only possible to predict the cycle time distribution of the lots for a few scenarios. In this paper, we present some model improvements which lead to a rather good cycle time prediction for a variety of load situations.

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.

In the next sections, we will outline some modeling approaches and results predicting cycle times 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 was not adequate. Then, we will present the improvement approaches. In (Rose 2007), we discussed some first ideas which are further developed in the material that is presented here.

2 PREDICTING CYCLE TIMES

Due to the success of our first studies, 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 (Figure 1).

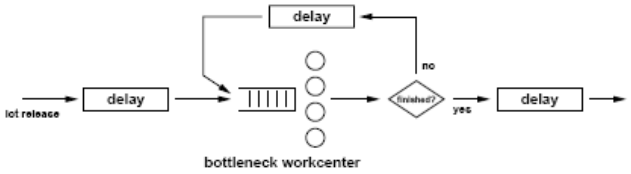

Figure 1: Simple Fab model

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 (Figure. 2).

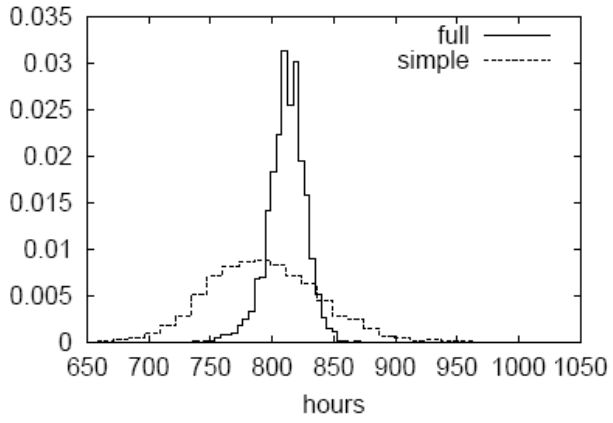

Figure2: Product cycle time distribution deviations

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. Figure 3 and 4 show this phenomenon. In Figure 3, the per-layer cycle times of the full model are clearly separated.

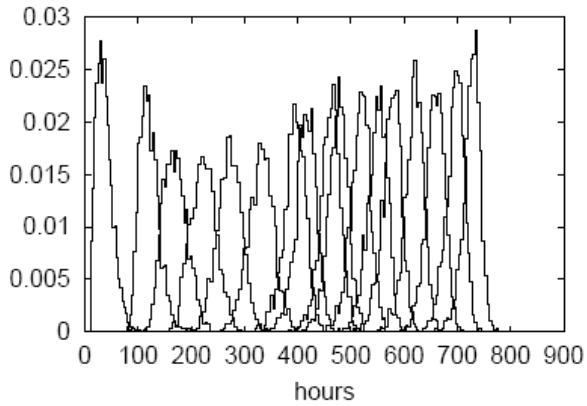

Figure 3: Per-layer cycle times of the full model

In Figure 4, however, the per-layer cycle time distributions become broader and broader from layer to layer.

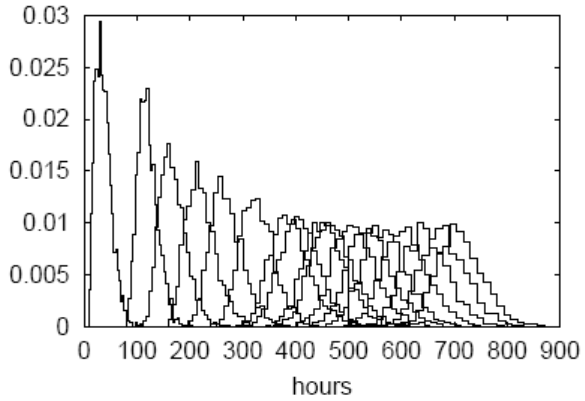

Figure:. Per-layer cycle times of the simple model

In addition to this weakness, the simple model of our earlier papers 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 Figure 5, 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. The model was calibrated for 85% utilization. At this load, it matches well the flow factor of the full detail fab model. For smaller utilization values, it overestimates the true flow factor and for larger ones the flow factors are underestimated considerably.

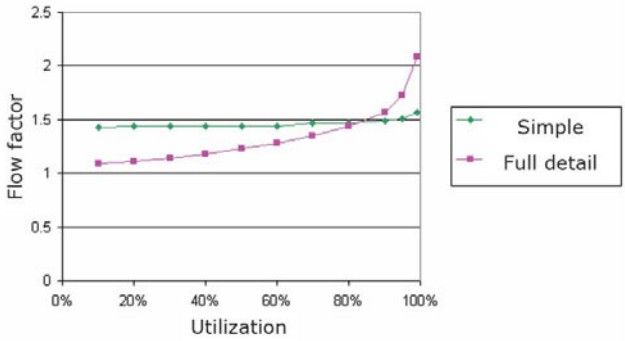

Figure 5: Comparison of characteristic curves

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

3 UTILIZATION-DEPENDENT BEHAVIOR

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 Figure6.

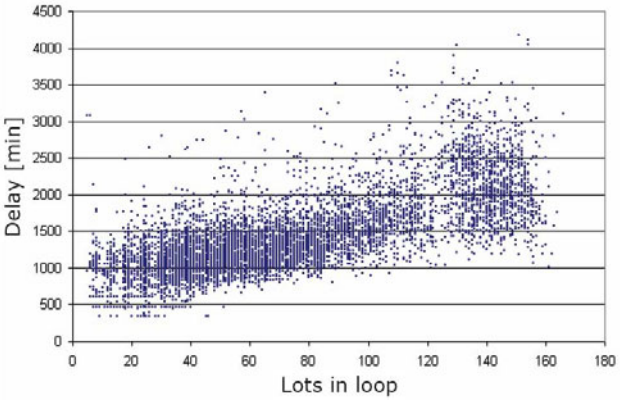

Figure 6: Inventory dependent delays

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 (Figure 7).

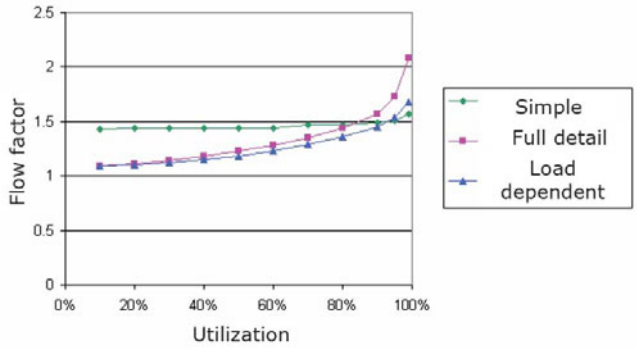

Figure 7: Improved characteristic curve

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

4 CYCLE TIME ESTIMATION

The main reason for this behavior is still the one mentioned in Section 2. As long as lots can pass each other while being delayed in the delay units, the cycle time modeling performance will remain inadequate.

As a consequence, we had to develop a new approach for delaying lots during the phase from leaving the bottleneck until re-entering it again. During this phase lot passing/overtaking should not be possible. The simplest approach to avoid passing is to put all lots into a queue which is served in FIFO order. In this case, however, the delay depends on the service time distribution of the processing unit working on the lots in the queue. We determined the service time distribution simply by computing the interarrival times of the lots at the bottleneck workcenter.

We used three ways to compute these values:

- Version 1: We use a single distribution for all lots (Figure 8).
- Version 2: We use individual distributions for all product changes, i.e., we need the square of the number of products of distributions (Figure 9).
- Version 3: We use individual distributions for all products (Figure 10). In this case, we need a queue for each product.

In all three cases, the service time distribution(s) of the processing unit serving the queue(s) depend on the load situation, i.e., we keep the dynamic behavior of model outlined in Section 3. In all three, cases the flow factor over the utilization is similar to the results presented in Figure 5.

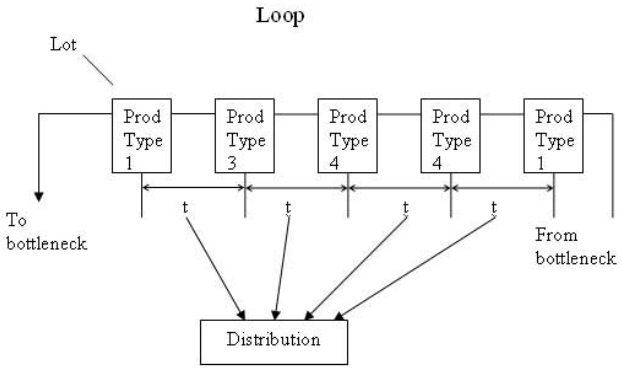

Figure 8: Estimation of the interarrival time (v1)

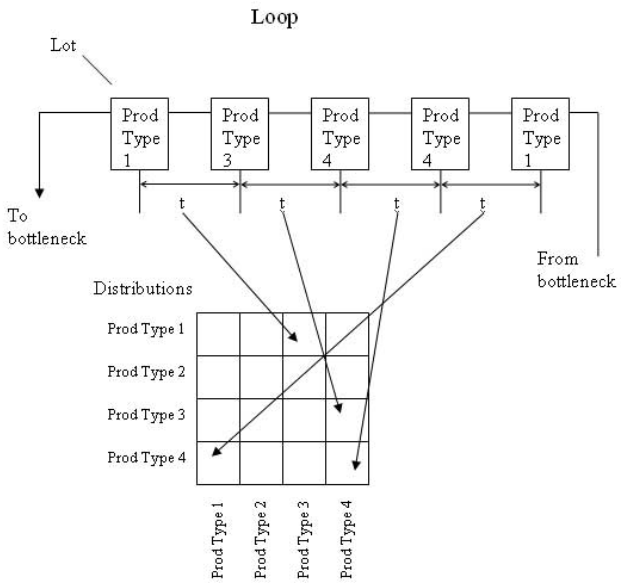

Figure 9: Estimation of the interarrival time (v2)

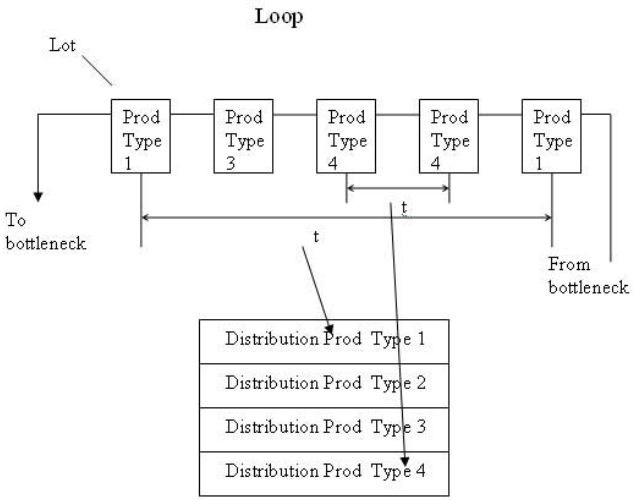

Figure 10: Estimation of the interarrival time (v3)

Figure 11 depicts the flow factor densities for the full detail model and the three versions of the simple model at a utilization of 90%. For version 1 and 2 the densities almost match and their shapes are very similar to the one of the full detail case. The average flow factors of these simple models is lower than for the reference case. For version three however the average is a little bit larger than for the full detail model and the density is to broad. The main reason for this behavior is that due to the usage of parallel queues lots of different products can pass each other while lots of the same products keep their sequence. Thus, overtaking is reduced but not completely avoided.

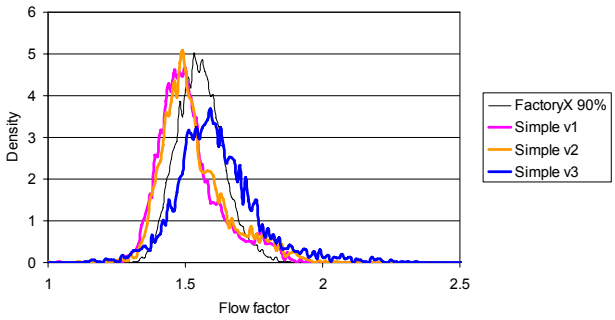

Figure 11: Flow factor densities

With respect to the cycle time density functions, the situation is different (Figure 12). In this case, version 3 clearly outperforms versions 1 and 2. It matches the full detail fab density very well.

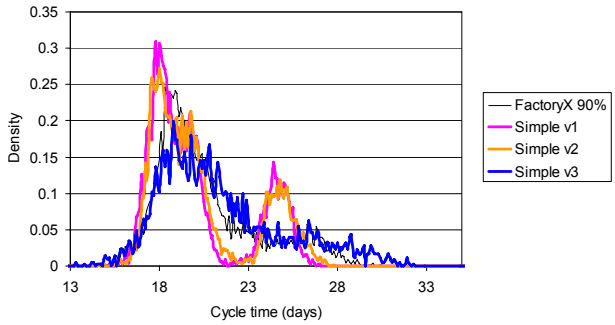

Figure 12: Cycle time densities

5 SUMMARY

In recent experiments, we were able to find a very simple model that mimics the load dependent behavior of a full detail model. It was not possible, however, to find good cycle time estimates for the full fab model with this simple model.

We presented a modification of the simple model that solves the cycle distribution problem to a large extend. Our approach is to replace the center delay unit by a single server with load dependent service times. We suggest three versions of this approach which differ in the way the service time distribution of the single server is estimated. The simulation results show that the cycle time estimates are considerable better compared to the earlier simple modeling approaches but that there is still room for improvement.

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

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

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

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

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

Rose, O. 2007. Improving the accuracy of simple simulation Models for complex production systems. In *Proceedings of the 2007 INFORMS Simulation Society Research Workshop*, Fontainebleau, July 5-7, 2007.

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>.