WIP EVOLUTION OF A SEMICONDUCTOR FACTORY AFTER A BOTTLENECK WORKCENTER BREAKDOWN

Oliver Rose

Institute of Computer Science
University of Würzburg
97074 Würzburg, GERMANY

ABSTRACT

In semiconductor fabrication facilities, an increase in work in progress (WIP)can be observed even weeks after the failure of the bottleneck workcenter. In this paper, we develop a simple fab model that facilitates the study of this phenomenon. The simplified factory consists of a detailed model of the bottleneck workcenter and a delay unit that represents the rest of factory. After passing the delay unit the lots are fed back to the bottleneck to model the cyclic flow of lots of real wafer fabs. We study the behavior of this model for numerous scenarios, and it turns out that by means of the model the WIP increase phenomenon can successfully be reproduced. In addition, we provide first results on how to avoid the unwanted increase in inventory.

1 INTRODUCTION

To assess the influence of various dispatching rules on wafer fab performance measures, simulation is used in general. 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 workcenter, i.e., all machines of the work center that constrains the fab capacity are down for a few days. This particular fab behavior was reported by fab managers of a Siemens memory fab.

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(Lovegrove, Hammond, and Tipper 1990). 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 in 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. We carry out several experiments that show how the simulation results change due to modifications in the model and in the dispatching rules. It turns out that the model is capable of reproducing the building up of inventory after catastrophic failure. In addition, we show that due-date based dispatch rules such as critical ratio lead to a worse fab performance in terms of WIP level and cycle time variations than FIFO dispatching for the period after the catastrophic failure. This is very much in contrast to the steady-state results where due-date based dispatching clearly outperforms FIFO dispatching (Brown, Fowler, Gold, and Schömig 1997).

The paper is organized as follows. The reduced factory model and the simulation details are presented in Section 2. Since the delay unit is a key part of the factory model, the effect of different delay models on the simulation results are provided in Section 3. In Section 4 we outline approaches to avoid the WIP increase after the catastrophic failure.

 The definitive version of this paper has been published in Winter Simulation Conference '98, 1998, 10.1109\/wsc.1998.745805.

2 FACTORY MODEL

Typical wafer fabs consist of several hundred machines producing tens of different products at a time. The wafers are manufactured according to recipes that contain several hundred processing steps. Due to the layered nature of semiconductors, the wafers visit sequences of machines several times, i.e. they are proceeding through the fab in cycles. Memory chips may have up to 30 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. The consequences of this permanent reordering of lot priorities will become apparent in Section 3.

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. Figure 1 shows the proposed factory model. It consists of a bottleneck workcenter, a delay unit, and a control unit. The bottleneck workcenter 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 workcenter 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 workcenter.

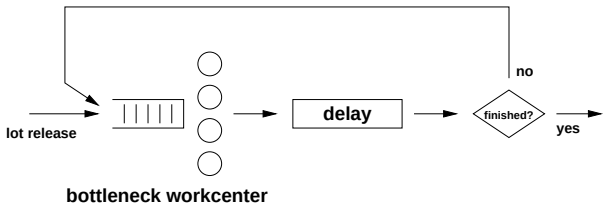

Figure 1: Factory Model

Simulation Details

For the simulation experiments we considered the following parameters. These parameters are not chosen arbitrarily but in conformance with current wafer fabs, e.g. as reported by Siemens fab managers.

There are four products that are manufactured by the fab. Each product has a lot start rate of 0.0925 lots/hour where the time between lot starts is constant. The processing times at the bottleneck workcenter are 0.7, 0.9, 1.1, and 1.3 hours, respectively. The processing times are assumed to be identical for each layer. All products have 10 layers. This results in a bottleneck workcenter load of 92.5 % ($= 4[\text{products}] \cdot 0.0925[\text{lots/hour}] \cdot (0.25 \cdot (0.7 + 0.9 + 1.1 + 1.3)[\text{hours}] \cdot 10[\text{cycles}]/4[\text{machines}]$). This is a reasonable load for the bottleneck workcenter if the average percentage of downtimes is assumed to be less than 7.5 %.

The average period of time spent in the delay unit is 125 hours for each product and layer. To facilitate the application of due date based dispatch rules, we require the processing times of all machines. Hence, we have to partition the delay time into a constant amount of 50 hours of processing time and a random amount of time with an average of 75 hours for waiting and other non-processing times. Details on the distributions of the waiting time are given in the next section. We set the lead flow factor to 2.5, i.e. the ratio of cycle time and raw processing time is intended to be 2.5. For the bottleneck tool this is to be achieved by adequate dispatching, whereas the rest of the fab represented by the delay unit is always conforming to the intended flow factor due to defining an average delay time of 125 hours and a processing time of 50 hours. At lot start, a due date of current time+ $10 \cdot$(125 hours+$2.5 \cdot$bottleneck processing time) is assigned to each lot.

The bottleneck workcenter consists of four machines. We consider the following four dispatch rules.

FIFO (First In First Out) The waiting lots are scheduled in the order of their arrival. This rule is the only one considered that does not lead to a reordering of queued lots.

SPTF (Shortest Processing Time First) The lots are scheduled according to their processing times. Lots with the shortest processing time are taken from the queue first.

CR (Critical Ratio) Each time a lot has to be taken from the queue, the following index is assigned to each of the waiting lots:

$$\text{CR} = \frac{\text{due date} - \text{current time}}{\text{remaining processing time}}.$$

The lot with the smallest index value is chosen for processing. As a consequence, lots that are closer to their due dates are preferred.

ST (Slack Time) Compared to CR, the index used for scheduling the lots is based on a difference

and not on a ratio:

$$\text{ST} = \text{due date} - \text{current time} - \\ - \text{remaining processing time}.$$

Again, the lot with smallest index is removed from the queue. The ST rule does not increase the priorities as fast as CR when lots are about to miss their due dates.

In the latter three cases, ties are broken by the FIFO rule. In contrast to the first two rules, the latter two rules take into account the due dates of the lots.

The above factory model is used below to determine the fab performance in time after a complete bottleneck workcenter breakdown. The simulation model was implemented in ARENA (Kelton, Sadowski, and Sadowski 1997), and the statistical postprocessing algorithms were developed by ourselves.

As a first step, we determine empirically the start-up phase of the system (Law and Kelton 1991). Having started with an empty system, the estimated end of the transient phase is at about 1500 hours.

Hence, we schedule the breakdown at 2000 hours of simulated time. All machines of the bottleneck workcenter become unavailable for processing. After 50 hours of repair time all machines start processing again. The simulation ends after 5000 hours of fab time.

During each simulation replication of 5000 hours, we record WIP changes and cycle times of finished lots. To reduce the amount of data and to facilitate the synchronization of the measurements from different replications, we apply the following method. The simulated time is divided into 10-hour intervals. For each 10-hour interval, we compute the sample mean of the cycle times of the lots that leave the fab during this period. With respect to the WIP, we compute the time-based average of the WIP level during each 10-hour interval, i.e., each WIP level observed during this period is weighted by the percentage of time during which it is kept. For each replication, we obtain a condensed WIP and cycle time sequence of 500 values each (5000 hours/10 hours).

To obtain statistically useful results, each experiment is repeated 500 times. The curves shown in the rest of the paper are based on averaging the condensed WIP and cycle time sequences of 500 simulation replications. The 95% confidence intervals of the WIP sequence of the fab with FIFO dispatching are shown in Figure 2. All other experiments lead to approximately the same ratios of confidence interval half-widths and sample means.

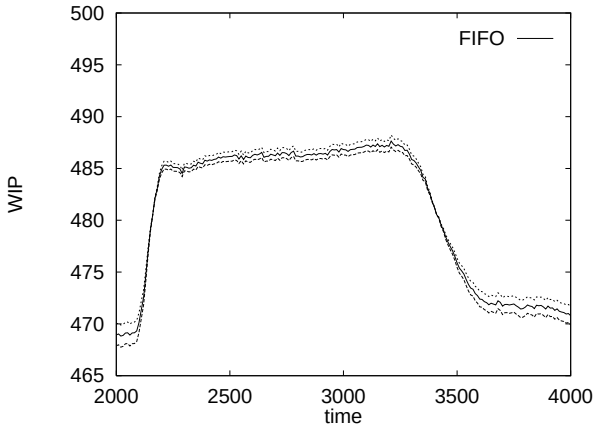

Figure 2: 95% Confidence Intervals of the WIP Sequence

3 MODELING OF THE DELAY

We begin our study with a set of experiments where we intend to determine the effect the delay time model on the behavior of the fab model. First, we consider a delay unit where the delay time is constant, i.e. 125 hours for all products and all layers.

Figure 3 shows the WIP evolution for constant delay under the regime of the four considered dispatching rules at the bottleneck workstation. In order to be able to show some interesting effects, the run length of these replications was extended to 8000 hours.

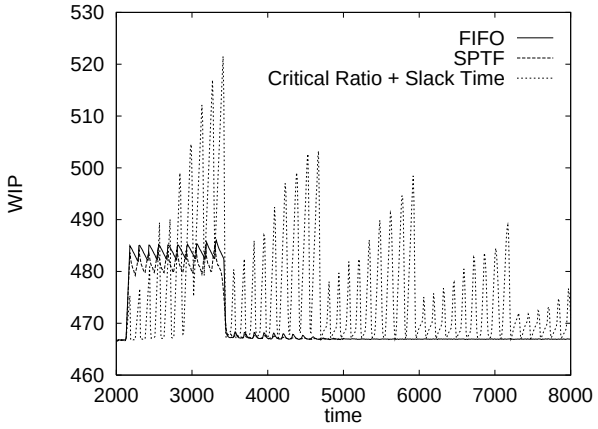

Figure 3: WIP Evolution for Constant Delay.

The two dispatch rules that do not consider due dates, FIFO and SPTF, show a fundamentally different behavior from CR and ST. During the breakdown the WIP increases very fast for FIFO and SPTF, stays approximately constant for about 1250 hours, and then drops down to the steady state level immediately. In the case of SPTF, the WIP level is lower

than for FIFO. After about 3500 hours, no effects from the breakdown can be observed in the fab. For CR and ST, however, the situation is different. WIP builds up more slowly, but about 500 hours after the breakdown it is becoming significantly higher than the FIFO level. Even after all lots that experienced the catastrophic failure left the fab, this behavior is repeated with decreasing WIP level. It is worth noting that there is almost no difference in the behavior of CR and ST for an intended flow factor of 2.5.

In all cases, the WIP level oscillates at a frequency of about $1/125[\text{hour}^{-1}]$. The reason for this oscillations is the constant delay introduced by the feedback loop to the bottleneck workstation. The lots waiting in the bottleneck center queue are processed very fast compared to the feedback delay and show up almost at the same time at the queue again.

Now, we explain the reason for the unexpected behavior of the system if due-date dependent rules are applied. Right after finishing the repair of the bottleneck machines, those lots are preferred by CR that are closest to their due date. Since all lots that saw the bottleneck down have a due date that is closer than all lots that are newly arriving after the end of the failure, all of the blocked lots are always processed ahead of the new lots. New lots will not be able to seize a server until all blocked lots left the queue. Therefore the WIP is building up due to the permanent arrival of new lots. As soon as all lots that experienced the failure have left the fab, the new lots that were blocked by these lots take their role and the phenomenon of increasing WIP repeats itself. Since the bottleneck workcenter has a spare capacity of 7.5 % the peak level of the WIP is becoming smaller and smaller.

Looking only at the WIP graphs, it is likely to draw the conclusion that SPTF is a reasonable dispatch rule in case of a catastrophic failure since it leads to the smallest WIP level and little WIP level variations. With respect to cycle times, this is no longer the case.

Figure 4 depicts the cycle time evolution for FIFO and SPTF dispatching.

While FIFO cycle times show the same behavior as FIFO WIP levels, the SPTF cycle times show periods where the cycle times are lower than the FIFO ones but also periods where the cycle times are considerably higher. The reason is the priorization of lots with smaller processing times at the bottleneck workcenter. Thus, the lots with a processing time of 0.7 hours rush through the bottleneck but lots with a processing time of 1.3 hours have to wait until lots of all other products have been processed.

Considering both WIP level and cycle times, there is a clear indication that FIFO dispatching will help

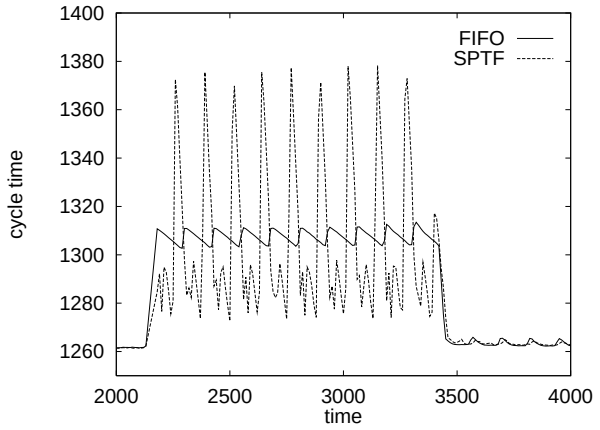

Figure 4: Cycle Time Evolution for Constant Delay

to avoid WIP build up and will considerably reduce the cycle time variations.

Up to this point we discussed a fab that is completely deterministic. Of course, a real fab is not. In particular, the waiting times experienced by a lot are far from being constant. Hence, we considered a new delay model that offers larger variations in the delay times. We chose an exponential distribution with mean 75 hours shifted by 50 hours, i.e. we assume a constant sum of processing times and an exponentially distributed sum of waiting times and other non-processing times. The variation of this model is higher than that of the Siemens fab.

Figure 5 shows the WIP evolution for shifted exponential delay.

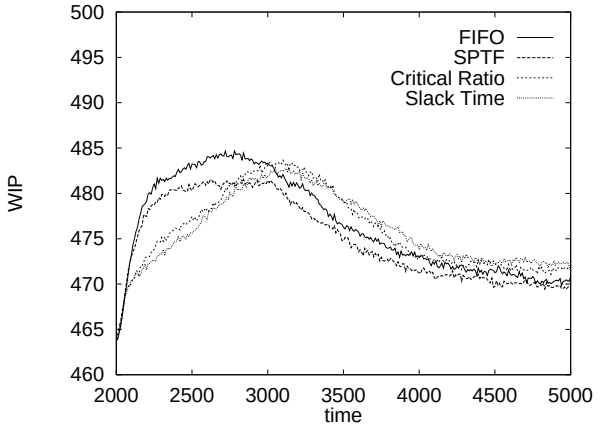

Figure 5: WIP Evolution for Shifted Exponential Delay

Neglecting the oscillations, the WIP evolution is similar to that presented for constant delays (cf. Figure 3). For FIFO and SPTF, the WIP stays on approximately the same level for about 1250 hours and

goes down to the steady-state level, whereas for CR and ST the WIP is constantly increasing during that period. The WIP decrease is considerably slower than for the constant delay fab. The oscillations disappear since the lots are no longer synchronized because they experience random delays due to the shifted exponential delay unit.

The two modeling approaches of constant delays and shifted exponential delays are the two extremal cases of delay time variability that were taken into consideration. In the following, we apply a delay model that is closer to real fab behavior: an Erlang-5 distributed delay time with a mean of 75 hours shifted by 50 hours. For details on exponential and Erlang distributions, see textbooks on simulation, e.g. (Law and Kelton 1991).

Figure 6 shows the WIP evolution for shifted Erlang delay.

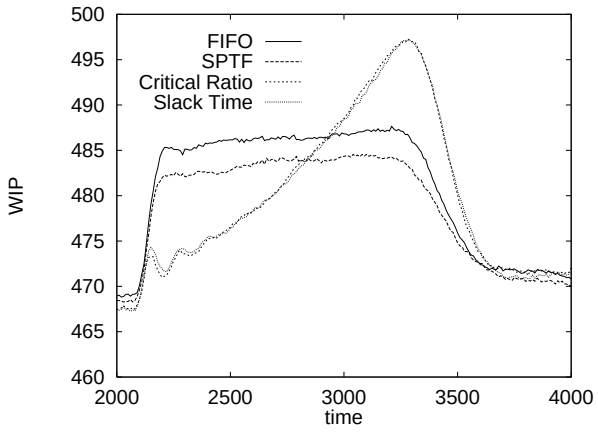

Figure 6: WIP Evolution for Shifted Erlang Delay

The observed behavior is a mixture of the constant delay scenario and the shifted exponential scenario. Apart from a few oscillations right after the repair of the bottleneck machines, the oscillation disappear due to the randomness of the delay time. Again, FIFO and SPTF dispatching lead to approximately constant WIP levels, and CR and ST to constantly increasing WIP levels.

With respect to cycle times, the shifted Erlang system, shows the behavior presented in Figure 7.

As for the constant case, the FIFO, CR, and ST cycle time sequences are approximately directly proportional to the respective WIP levels. In contrast to the constant case SPTF cycle time graph, the shifted Erlang one has no peaks higher than the FIFO curve. Due to the randomness in the delay times, the order of the lots is somewhat rearranged during the passage of the delay unit. Thus, there are not always the same lots competing for service at the bottleneck

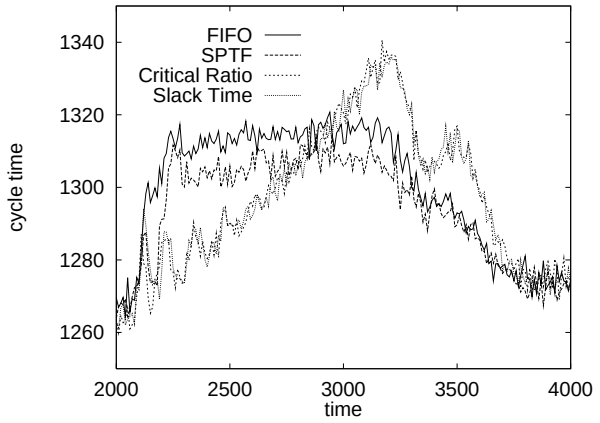

Figure 7: Cycle Time Evolution for Shifted Exponential Delay

workcenter.

Table 1 shows the average and variance of the cycle times observed in the time interval from 2000 hours to 4000 hours. With respect to the average cycle times,

Table 1: Cycle Times from 2000 Hours to 4000 Hours

Dispatch rule	Average	Variance
FIFO	1300.0	284.6
SPTF	1294.2	210.0
Critical Ratio	1297.0	409.1
Slack Time	1296.8	388.0

all four dispatch rules provide the same results. The variances of cycle times, however, are considerably different. The due-date based dispatch rules lead to a variance in cycle times that is about twice as large as the SPTF variance. The FIFO variance lies between the SPTF and Slack Time ones.

From the results of our experiments, we draw the following conclusions. After a catastrophic failure of the bottleneck workstation, the reduced fab model introduced in Section 2 shows essentially the same behavior as reported from a real wafer fab. Under the regime of CR dispatch WIP level and cycle times are increasing and reach their maxima several days to weeks after the end of repair. This behavior can be observed for delay unit models with different variability. In our experiments the variability ranged from none, i.e. constant delays, to shifted exponential. For small amounts of variability the WIP level tends to oscillate considerably. With respect to the average WIP level and cycle time, FIFO dispatch leads to about the same results as CR for the period from leaving the steady-state until returning to it. Under

the FIFO rule, however, less variations in WIP and cycle time are observed and the maxima of both measures are smaller. If the delay time variability is not too small, SPTF provides even better results.

With respect to explaining the WIP increase even weeks after a catastrophic failure, we conclude that this fab behavior is caused by the combination of due-date oriented dispatch and the cyclic nature of the flow of lots through the fab. Only these two typical characteristics of wafer fabrication together, lead to the blocking of fresh lots at the bottleneck workcenter and induce the constant WIP increase.

4 AVOIDING THE WIP INCREASE

Since unnecessary WIP is a waste of money and cycle times that are longer and more variable than expected are causing trouble, fab managers try to avoid such situations.

For our catastrophic failure scenario, we will not be able to apply complex strategies due to the simplicity of the model and the lack of parameters to play with. In the following, we present two simple strategies: changing of the dispatch rule and stopping lot release during repair time.

The strategy of changing dispatch rules to avoid WIP increase is based on the following observation. For the first period of time after the repair of the bottleneck tool, CR outperforms FIFO with respect to WIP level. During the second period, however, FIFO leads to smaller amounts of inventory than CR (cf. Figure 6). Unfortunately, as shown in Figure 8, this strategy does not work.

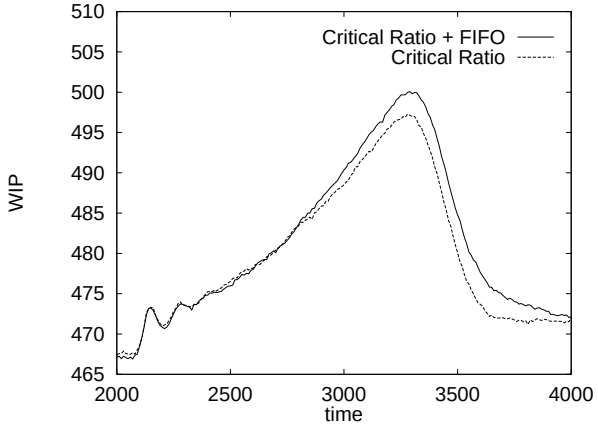

Figure 8: WIP Evolution for Changing Dispatch from CR to FIFO

Even though the dispatch rule changes from CR to FIFO at time 2700 hours WIP is increasing as if the rule would have been left unchanged. This indicates that the reordering of the lots that takes place after the repair determines the evolution of the WIP to a large extend. Later changes have only a minor influence.

As a second strategy, we tried an approach suggested by Goldratt (Goldratt and Cox 1994). As soon as the bottleneck workcenter of a fab that is needed for the processing of each product goes down the fab has zero capacity. Thus, it makes no sense to release new material into the fab since it will be blocked. The lot release should be restarted as soon as the bottleneck tool group is up again. Figure 9 shows the the WIP graphs for a fab where lot release was stopped from 2000 hours to 2050 hours.

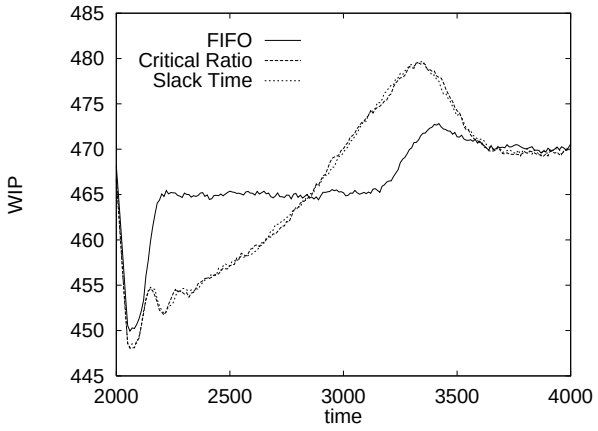

Figure 9: WIP Evolution for Stopping Lot Release

During the repair time of 50 hours, the WIP level drops considerably. Later on, the WIP evolution is identical to the conventional system shifted by the WIP drop. For FIFO, the maximum WIP level is only slightly higher than steady-state. For CR and ST, the maximum WIP level is higher than for FIFO but considerably lower than for the original system. With respect to the cycle times, the stop of lot releases has only marginal effects. In summary, the stop strategy provides a clear benefit with respect to WIP level but no effect with respect to cycle times. In addition, one has to take into account that there must be some spare capacity at the bottleneck because the lots that were not released to the fab during repair time will have to be released to it later on.

5 CONCLUSION AND OUTLOOK

In this paper, we present a reduced wafer fab model that exhibits essential features of a real wafer fab. It consists of a detailed model of the bottleneck workcenter and a delay unit that models the remaining

machines of the fab. Lots released to the fab model have to cycle through bottleneck and delay unit repeatedly in order to model the layered nature of semiconductor manufacturing. For our experiments, four dispatch rules at the bottleneck workcenter are assumed. Two of them without due date dependence (First In First Out, Shortest Processing Time First), and two of them with due dates involved (Critical Ratio, Slack Time).

This fab model is used to assess the evolution of the WIP level and cycle time of the fab after recovering from a catastrophic failure, i.e., a complete failure of all bottleneck machines for a longer period of time.

Particular attention is devoted to the modeling of the delay time for the delay unit. It turns out, however, that the dispatch rules have a greater effect on the behavior of the fab than the choice of the delay time model.

Using the proposed model, we were able to reproduce fab behavior as observed in real semiconductor manufacturing facilities. It turns out that the phenomenon of increasing WIP is mainly caused by a combination of the due-date oriented dispatching and the cyclic nature of the lot flow. Using the simple strategy of stopping lot release during repair time, we provide first results on how to partially avoid the tremendous increase in inventory after a bottleneck breakdown.

Currently, we consider less catastrophic failures of bottleneck workcenters, i.e., not all machines have to be repaired at the same time. The results of this study may be used to develop maintenance plans for bottleneck tools that provide as less increase in WIP and cycle times as possible. In a further study, we use real fab data to determine the mandatory statistical properties of the delay times, such as distributions and correlations, that allow for a good prediction of fab behavior through the simple fab model.

ACKNOWLEDGEMENTS

The author would like to thank Holger Oehring for his valuable programming efforts, and Hermann Gold and Alexander Schömig (Siemens AG) for fruitful discussions by which we were able to keep the study both academic and useful for industry.

REFERENCES

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

Brown, S., J. Fowler, H. Gold, and A. Schömig (1997). Measurable improvements in cycle-time-constrained capacity. In *Proceedings of the 6th International Symposium on Semiconductor Manufacturing.*

Goldratt, E. M. and J. Cox (1994). *The Goal: A Process of Ongoing Improvement* (2nd ed.). North River Press.

Kelton, W. D., R. P. Sadowski, and D. A. Sadowski (1997). *Simulation with Arena.* New York: McGraw–Hill.

Law, A. M. and W. D. Kelton (1991). *Simulation Modeling & Analysis* (2nd ed.). New York: McGraw–Hill.

Lovegrove, W., J. Hammond, and D. Tipper (1990). Simulation methods for studying nonstationary behavior of computer networks. *IEEE Journal on Selected Areas in Communications 8*(9), 1696–1708.

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

AUTHOR BIOGRAPHY

OLIVER ROSE is an assistant professor in the Department of Computer Science at the University of Würzburg, Germany. He received an M.S. degree in applied mathematics and a Ph.D. degree in computer science from the same university. He has a strong background in the performance evaluation of high-speed communication networks. Currently, his research focuses on the analysis of semiconductor and car manufacturing facilities. He is a member of IEEE and INFORMS.