Universität Würzburg
Institut für Informatik
Research Report Series

Analysis of the Short Term Impact of Changes in Product Mix

M. Dümmler and O. Rose

Report Nr.: 254 März 2000

Lehrstuhl für Informatik III,
Universität Würzburg
Am Hubland, 97074 Würzburg
E-Mail: `duemmler@informatik.uni-wuerzburg.de`

Analysis of the Short Term Impact of Changes in Product Mix

Mathias Dümmler, Oliver Rose
Lehrstuhl für Informatik III, Universität Würzburg, Am Hubland, 97074 Würzburg
E-Mail: `duemmler@informatik.uni-wuerzburg.de`

Abstract

In this paper we use a simulation model of an existing semiconductor fab to study the effects of changes in product mix on fab performance. We observe how short-term increases in wafer starts of a product influence the cycle time and WIP of this product and of the other products. It is examined how the fab recovers from production surges under different dispatch rules and how fab performance is affected if the product mix is changed on a weekly basis. The methodology used to study the transient behavior of a wafer fab using simulation is discussed.

1 Introduction

In modern semiconductor fabs, up to several dozens of different product types with up to several hundreds of derivatives are processed in parallel. Product mix is subject to constant change, due, for example, to incoming orders. In fabs that mainly produce customer specific chips ("make to order") the product mix depends strongly on the current amount of orders. Furthermore, with process technologies advancing, start rates of old technologies are continuously reduced while start rates of new technologies are being increased, leading to permanent changes in product mix. Since a large number of the machines in a fab are shared by different products, there is a strong interaction among them. Therefore, product mix has considerable impact on throughput, cycle times and hence on the capability of meeting due dates, which is considered to be one of the most important metrics to measure fab performance.

A typical question that arises in production planning is what short term effects an increase in the number of wafer starts of a specific product will produce. Can the resulting cycle times be tolerated? Can the increase in WIP (work in process) be handled? Is the fab able to recover after a production surge, i.e., do the cycle times return to a "normal" level?

To address this question in a simulation study, the transient behavior of the fab model has to be considered. In (Rose 1998) this approach is taken to analyze the behavior of a semiconductor fab after a breakdown of the bottleneck workcenter. This kind of study differs from the majority of studies on semiconductor fabs that are published which investigate the steady state behavior of the model, i.e., the long term average of performance characteristics like cycle time or WIP. See (Janakiram and Morrison 1999) for an example.

The approach taken in this paper is similar to that in (McKiddie 1995) and (McKiddie et al. 1994). The authors perform simulation experiments of a number of different surge scenarios in a fab producing a single type of wafers. They use regression analysis to develop a model that predicts the impact of a production surge on the cycle time of the product. This regression model can provide an approximation of the actual fab performance. However, the authors mention that the model is fab specific and that the results can not be generalized.

In this paper we present first results of a series of simulation studies to analyze the effects of changing product mix on fab performance. The fab model was built using data of an existing semiconductor fab with multiple products.

The performance characteristics observed are cycle time, WIP, and the number of finished wafers. In the first part of the study we focus on the impact of short time increases of start rates (surges) of a specific product on these performance metrics. The behavior of the fab under different dispatch rules at the machines is determined and it is observed how the fab recovers from such surges. In the second part we consider the effect that a frequently (in our case: weekly) changing product mix has on fab performance.

This paper is organized as follows. The simulation model of the semiconductor fab under investigation is presented in the following section, along with a definition of the dispatch rules applied in the study. The methodology used to derive performance characteristics from the simulation output is presented. Several of the results produced so far are presented and discussed. We conclude the paper with a summary and some directions for future research on this topic.

2 Simulation Model

The simulation model for this study has been developed using data from an existing wafer fab producing both logic (six different products) and memory (four different products) ICs. The fab model consists of some 600 machines and 120 operators. Lots are started in constant intervals, with different start rates for the ten products. Since some of the products require a batch operation as the first processing step, they are released in groups of three lots. The other products are released as single lots. A total of approximately 10,000 wafers are started per week in the original model, resulting in an utilization of the bottleneck workcenter of 90%. The machines are subject to downtimes due to inspections and failures. Other features of the actual fab like sequence dependent setup times, scrap, rework, and lot transportation are also part of the simulation model.

Dynamic dispatching is used at each machine to decide which lot in queue is to be processed next. In this study we compare three different dispatch rules:

- *Critical Ratio* (CR), which gives priority to the lot in queue with the lowest value of *Ratio* computed according to

$$\mathrm{Ratio} = \frac{\mathrm{Due\ Date - Current\ Time}}{\mathrm{Remaining\ Processing\ Time}},$$

- *First In First Out,* which prioritizes the lot that first entered the queue,
- *WorkAPD*, which uses the WorkStream® priority function (WPF) to choose a lot for processing. It is implemented as follows. The time *D* until the due date of a lot is computed according to

$$\mathrm{D} = \mathrm{Due\ Date - Current\ Time},$$

and the estimated remaining cycle time *C* is derived according to

$$C = \text{Lead Time Factor} \cdot \text{Remain. Proc. Time}.$$

The expected lateness *L* of the lot is

$$L = C - D.$$

Based on the values of L and D, the WPF is computed as follows:

$$\text{WPF} = \begin{cases} -100 \cdot \dfrac{L}{1+D} & \text{, if } L < 0 \\ -10 \cdot \dfrac{L}{1+D} & \text{, if } L \geq 0 \text{ and } D \geq 0 \\ -L \cdot (10-D) & \text{, if } D < 0 \end{cases}$$

If WPF is greater than 99.9 or less than –99.9, it is truncated to 99.9 or –99.9, respectively. The lot with the lowest value of WPF is processed first.

Unless mentioned otherwise, WorkAPD is used as dispatch rule.

3 Methodolgy

In order to investigate the short term impact of changes in product mix on performance characteristics we perform a transient analysis of the semiconductor fab using simulation. This kind of analysis is different from most of the common studies in that it does not take the long term behavior of the fab into account, but the evolution of performance characteristics over time.

The performance characteristics under investigation are the average cycle time of lots, the average amount of work in process, and the average number of finished wafers (wafer outs). To derive statistically significant results, ten independent replications were run for each simulation experiment. In this way, the inherent variability of the performance metrics, caused for example by random machine breakdowns, can be separated from the induced variability of the metrics caused by the changes in product mix (McKiddie 1995). All averages were taken over a period of one week. The average values of a particular week were then averaged over the ten replications. All models were run for a simulated time of three years.

To build the fab model and to perform the simulation experiments, the Factory Explorer™ simulation tool was used. The run time for one replication was about 45 minutes on a 266MHz Pentium II processor.

4 Results

In this section, we present some of the experiments performed so far. Two different types of scenarios are modeled. In the first type of scenario, the start rate for a single product is increased, creating a surge impulse which possibly leads to an overload situation of the fab. In the second type of scenario, start rates for all products are changed weekly to reflect the fluctuations in product mix caused by the production planning process.

4.1 Surge Analysis

In all of the following experiments the product mix is kept constant for one year. After that year a change in product mix is introduced by increasing the start rate of a single product. We will refer to this product as the "surge product". The length of the surge impulse is either one, three, or six months. The amount of lots that are started additionally during the surge is selected among the values 17%, 33%, 67%, and 100% of the original start rate. After the surge, the start rate is reset to the original value. The start rates of the other products are kept constant during the three years of fab operation.

Figure 1 displays a chart with graphs of the three performance metrics of the surge product. WIP (in wafers) and the amount of finished wafers are displayed using the same scale. The cycle time graph displays the average cycle time of the wafers that leave production in the corresponding week.

In this particular experiment the surge length is three months, and the additional amount of wafer starts of the surge product during the surge is 67% of the original start rate. The horizontal lines denote the begin (in week 53) and the end (in week 65) of the surge impulse.

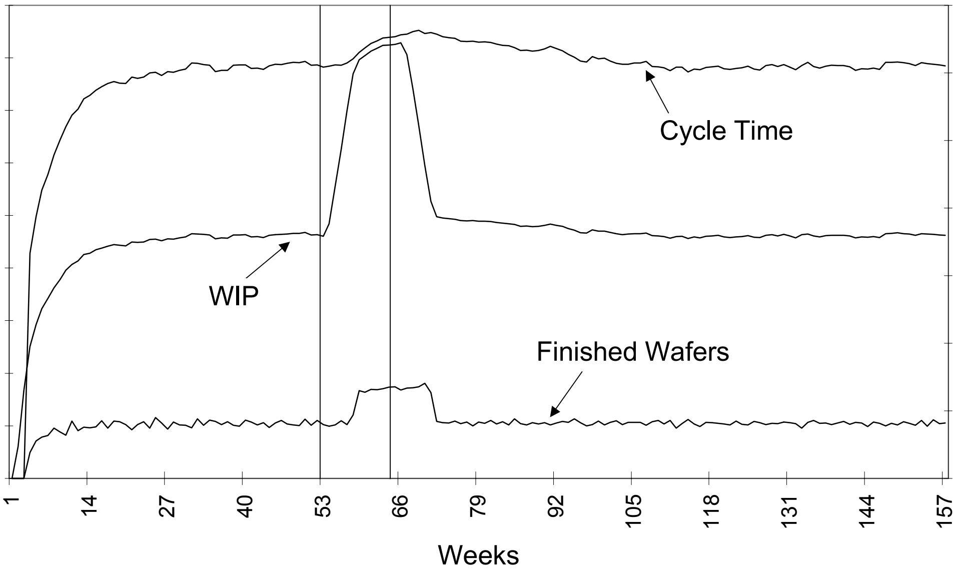

Figure 1: Fab performance in surge situation: Influence on cycle time, WIP, and wafer outs

After the surge impulse is initiated, a linear increase in WIP can be observed. The increase in WIP becomes less steep after a few weeks, coinciding with an abrupt increase in the number of finished wafers. The time offset to the increase in finished wafers corresponds to the cycle time of the wafers started at the beginning of the surge.

Additionally, the cycle time of the surge product is increased due to the increased start rate. Obviously the fab is able to recover from the surge, because after the surge

impulse the cycle time and the WIP level return to the level before the surge, although it takes significantly longer for the cycle time to return to the original level.
In Figure 2 we can observe the impact that different intensities of increasing the start rate have on the cycle time of the surge product. The idea in this experiment is to change the surge length and the surge height (amount of additional lots started) while keeping the product of surge length and height constant. For example, the curve "1 Month @ 200%" corresponds to the case were the start rate for the surge product is doubled for one month. Hence, in all three cases the amount of additional lots produced is the same, however the way the production surge is introduced differs. The vertical line at week 53 denotes the beginning of the surge impulse.
Since in the case where the surge impulse lasts only one month, it takes the longest for the cycle time to return to a normal level, one can conclude that it is advantageous to keep the increase in start rate as small as possible and to spread the additional lot starts over a longer time period instead. This conclusion was confirmed by similar experiments that were performed in the context of this study.

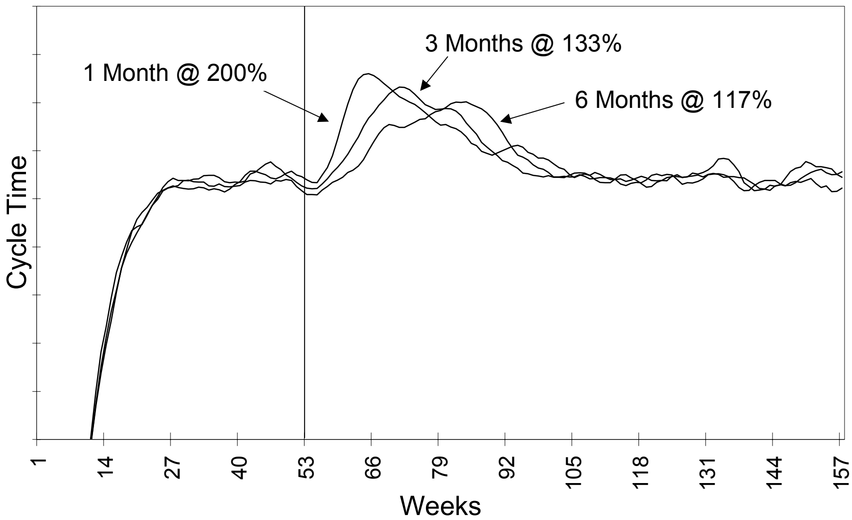

Figure 2: Impact of different surge lengths and sizes on cycle time

One aspect of this study was to investigate the behavior of the fab during a surge situation under different dispatch rules. Figure 3 shows the cycle time of the surge product during a surge with a length of three months. The start rate is incremented by 67% during the surge. The three dispatch rules compared are Critical Ratio, FIFO, and WorkAPD. For the Critical Ratio rule, a constant offset of 2.8 times the raw processing time (2.8XRPT) has been chosen to compute the due date of each lot.

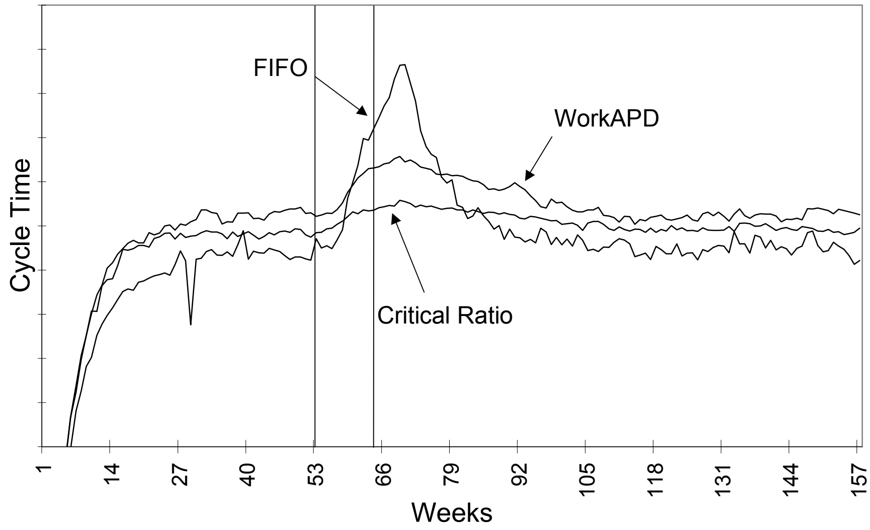

Figure 3: Cycle time under different dispatch rules

It is interesting to see that while the average cycle times before and after the surge are lowest for the FIFO rule, they become the highest during the surge for this rule. Cycle times are slightly higher under CR for the base product mix, but they increase only little during the surge. For the parameterization chosen for this experiment, WorkAPD performs worse than CR.

One can conclude that it is advantageous to change the dispatch rule from FIFO to CR during the surge phase. Previous studies implementing such an approach (Rose 1998) showed no improvement in cycle time. For the specific model of this study this result still has to be verified.

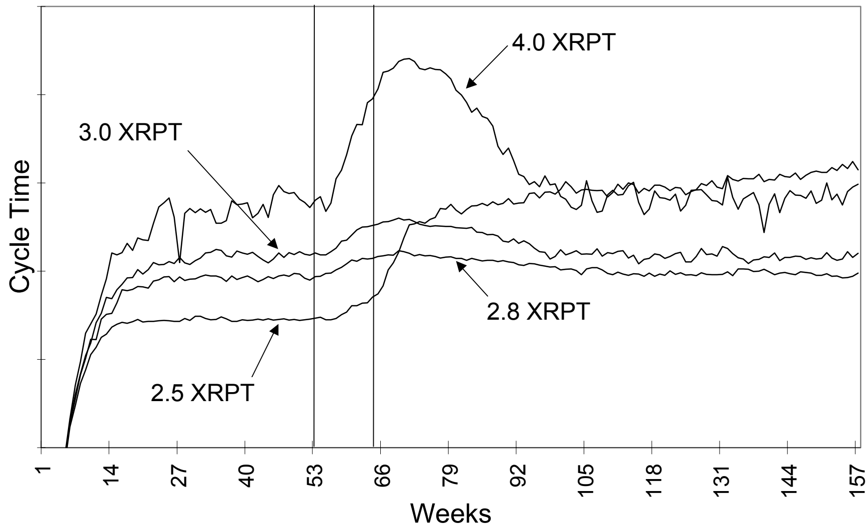

Figure 4: Parameterization of the Critical Ratio rule

The influence of the due date setting was investigated in another experiment. Figure 4 shows the cycle time curves for the same surge scenario as in the previous experiment. The four experiments were performed using the CR rule with different due date settings.

For a due date offset of 2.5 times the raw process time (2.5XRPT), the cycle times for the base mix in the first year of production are the lowest. However, during the surge phase the cycle times increase and keep doing so even after the surge period is finished. Hence, the fab is not able to recover from the surge. Setting the due date offsets to larger values (3XRPT and 4XRPT), leads to better performance during the surge, however cycle times are higher for the base product mix. By experimentation we were able to find an optimal value of 2.8XRPT for the due date offset which produced acceptable cycle times before and after the surge and which at the same time made the system very robust to production surges.

In Figure 5 we consider the surge impact on the work in process for the surge product and for the nine other products in a scenario where the start rate of the surge product is doubled for one month, and is then reset to the original value.

A clear peak of the WIP level can be identified for the surge product, with a steep rise shortly after the surge impulse begins and an equally steep decline shortly after the surge. When considering the other products, however, we observe a different behavior. The time offset from the beginning of the surge impulse to an increase in WIP for these products is larger for these products. It also takes longer, compared to the surge product, until the WIP levels return to their original values.

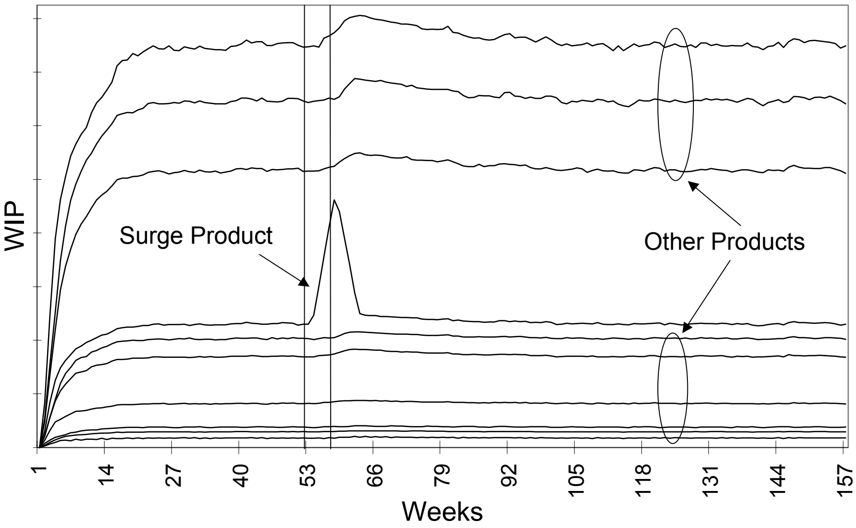

Figure 5: WIP levels

In contrast to the WIP evolution, the cycle times of the different products, depicted in Figure 6, exhibit a parallel evolution. The relative increase in cycle time is about 5% to 6% for the different products.

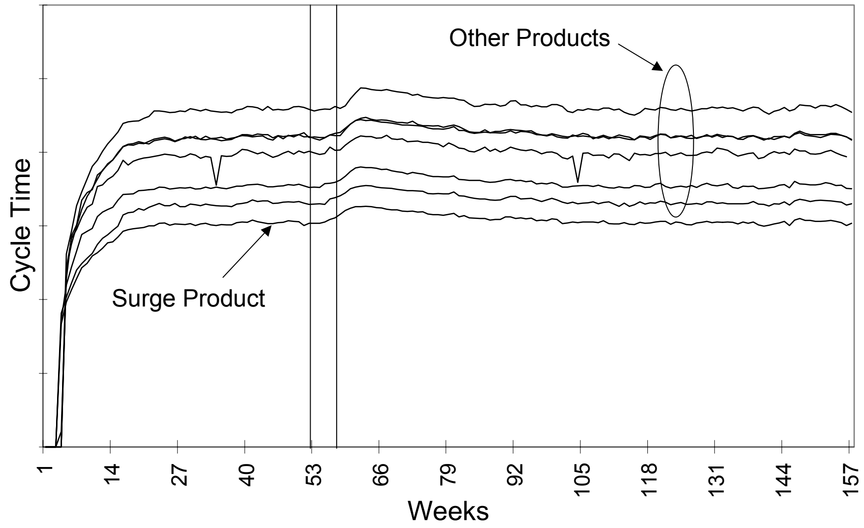

Figure 6: Cycle time of other products

Observing the evolution of the average number of finished wafers of the different products in Figure 7 reveals that the additional amount of wafers of the surge product finished during the surge impulse can only be maintained at the expense of the other products. This is explained by the fact that the surge impulse generates an overload situation at the bottleneck workcenter.

The most important conclusion drawn from Figure 7, at least in terms of production planning, is that the additional demand for wafers of the surge product can be satisfied. Hence, production planners can use this kind of simulation experiment to assess the effects that increasing the start rate due to incoming orders has on fab performance.

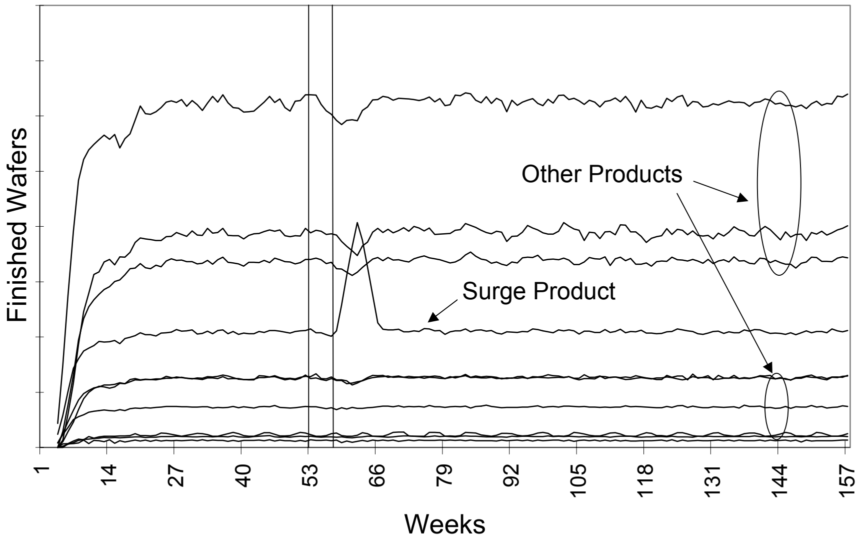

Figure 7: Finshed wafers

4.2 Weekly Changing Product Mix

The production plan for the wafer fab under investigation in this study is generated on a weekly basis, based on the incoming orders and the current fab capacity. To reflect the changes in product mix that result from this planning process, we generated a simulation model of the fab where start rates are computed anew for each week. In Figure 8, the cycle time curves of two products for this model and for a model with constant product mix are displayed.

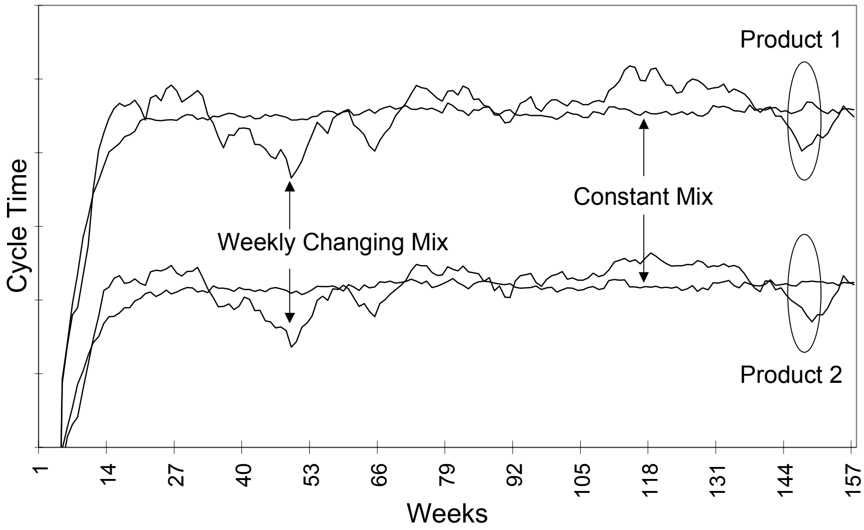

Figure 8: Constant vs. changing product mix

The target start rates of each product, i.e., the average number of lots started, are identical in both models. For the model with changing product mix the start rate for a given product for a particular week is generated by multiplying the target start rate of this product by a factor that is generated from an uniform distribution on the interval (0.7, 1.3). This leads to a maximum deviation of the start rate from the average start rate for a given week of up to 30%.

The average cycle time observed is identical in both models. The model with changing product mix exhibits cycle times that deviate up to 13% from the average cycle time. Compared to the maximum deviation of 30% in product mix this might be an acceptable cycle time performance. However, considering the average number of tardy lots, i.e., the number of lots that leave the fab after their due date, reveals the problems arising from fluctuations in product mix: While for the constant mix only 2% of the lots are tardy on average, 16% of the lots can not be finished before their due date for the changing product mix case.

5 Summary and Conclusion

This paper presented some results of a series of studies performed to analyze the impact that changes in product mix have on fab performance. We considered two types of changes, namely a surge scenario where the start rate of a single product is increased for a certain period of time, and a scenario where the start rates for each product are changed at the beginning of each week.

Experiments with different dispatch rules showed that the choice of a specific rule has a significant impact on how the fab can handle changes in product mix and short term overload situations. We also found that the due date setting, i.e., the appointment of the date when a lot should be finished processing, affects the short term as well as the long term behavior of the fab if due date based dispatch rules like Critical Ratio and WorkAPD are applied.

Future research will include an investigation whether the results gained for the particular fab model used in this study can be generalized to other types of wafer fabs. Currently, investigations in lot start mechanisms that allow to alleviate the effects of sudden changes in product mix are performed.

References

Janakiram, M., and Morrison, J. 1999. Capacity Planning and Study of Scheduling Policies Using Simulation at Motorola's ACT Fab. *Proceedings of the 1999 International Conference on Semiconductor Manufacturing Operational Modeling and Simulation,* 201-206.

McKiddie, R. 1995. Some 'No-Panic' Help for Wafer-Start Surges. *Semiconductor International,* 115-120.

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. *SEMATECH Technology Transfer 94022223A-XFR..*

Rose, O. 1998. WIP Evolution of a Semiconductor Factory after a Bottleneck Workcenter Breakdown. *Proceedings of the 1998 Winter Simulation Conference,* 997-1003.