Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

ANALYSIS OF THE INSTATIONARY BEHAVIOR OF A WAFER FAB DURING PRODUCT MIX CHANGES

Mathias A. Dümmler

Department of Distributed Systems
Institute of Computer Science
University of Würzburg
Am Hubland, 97074 Würzburg, GERMANY

ABSTRACT

This paper presents a series of experiments that were conducted to investigate in the instationary behavior of a wafer fab after changes in product mix. The experiments were performed using a simulation model of the front end area of an existing semiconductor fab. 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 such production surges under different dispatch rules. We also investigate how different lot start mechanisms affect the short term fab performance. More specifically, we observe the effects of changing the mix of the lots started into the fab on a weekly basis. Finally, we compare two alternative ways of releasing lots into the fab: The first way is to distribute lot starts evenly over a given period, e.g. a week, the other way is to start all lots at the beginning of the period.

1 INTRODUCTION

In a recent paper (Dümmler and Rose 2000) we investigated in the effects that yield from short term changes in product mix in a semiconductor fab. Some of the results presented there will be reviewed in this article. We also provide new results from current research.

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, for example due 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 competition for resources. 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 these questions in a simulation study, the transient behavior of the fab model has to be considered. Rose (1998) uses this approach 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 performance characteristics observed in this paper are cycle time, WIP, and the number of finished wafers (wafer outs). 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. Finally, we compare to ways of releasing raw material, i.e., wafers, into the fab. The first way is to distribute the releases evenly over a given period, e.g. a week, the second way is to release all lots at the beginning of each week.

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. 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

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

- First In First Out (FIFO), 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

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

and the estimated remaining cycle time C is derived according to

$$C = \text{Lead Time Factor} \cdot \text{Remaining Processing Time}.$$

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:

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

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, 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. Three 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. In the third scenario, two ways of releasing lots into the fab are compared.

Please note that due to confidentiality reasons no absolute values can be provided in the cycle time and WIP graphs and in the tables.

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 aver-age 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.

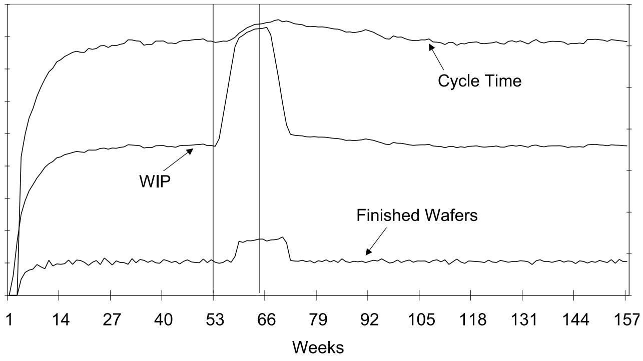

Figure 1: Influence on Cycle Time, WIP, and Wafer Outs

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 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 per day) while keeping the total amount of additional lots started constant. For example, the curve "1 Month @ 200%" corresponds to the case were the start rate for the surge product is two times the original value during 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.

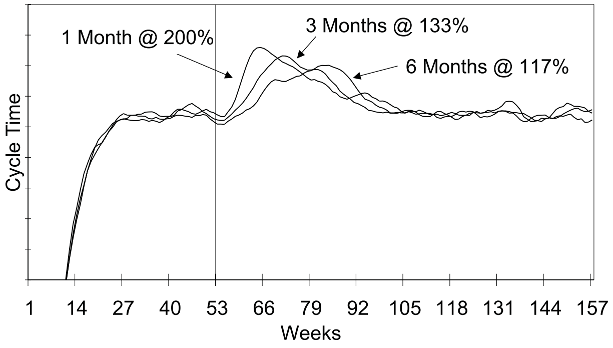

Figure 2: Impact of Different Surge Lengths and Sizes on Cycle Time

In the case where the surge impulse lasts only one month, it takes longest for the cycle time to return to a normal level, thus 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 experiments of this study.

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 increased 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) was chosen to compute the due date of each lot.

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 largest during the surge for this rule. Cycle times are slightly larger 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.

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.

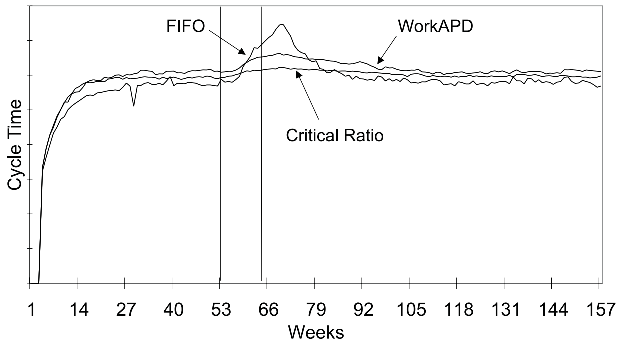

Figure 3: Cycle Time under Different Dispatch Rules

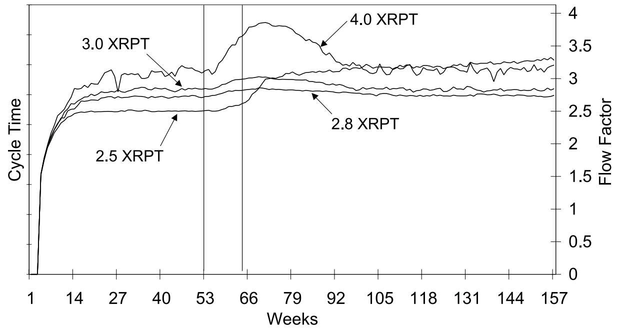

Figure 4: Parameterization of the Critical Ratio Rule

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 larger for the base product mix. By experimentation, we were able to find an optimal value of 2.8XRPT for the due date offset that produced acceptable cycle times before and after the surge and which at the same time made the system very robust to production surges.

To examine whether the desired due date offset, given by the value of XRPT, is actually achieved, we also observed the flow factor of the surge product. The flow factor, or X-theoretical value, is defined as the fraction of the total cycle time of a lot and its raw processing time without delays. Since the flow factors can be derived from the cycle times by dividing by a constant factor, the flow factor graph is the same as the graph of the cycle times.

From Figure 3, one can conclude that it might be advantageous to change the dispatch rule from FIFO to CR during the surge phase. Therefore, we performed experiments where the FIFO rule was used during the non–surge phase and during the surge impulse we switched to CR. Figure 5 presents the results of these experiments. The surge impulse had a length of one month and the start rate of the surge product was doubled during this month. Note, that Figure 5 zooms into the interesting part of the graph (weeks 47 to 77).

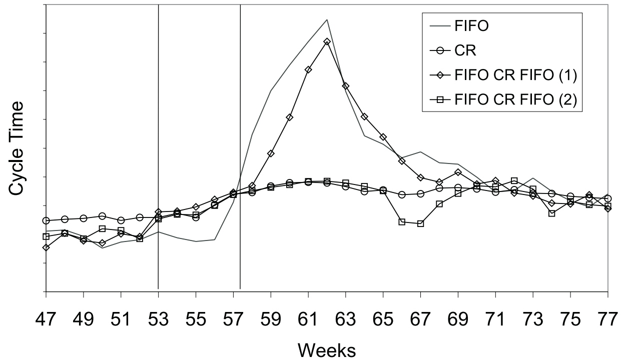

Figure 5: Changing Dispatch Rule During Surge

The curve labeled “FIFO CR FIFO (1)” shows the cycle time evolution if the FIFO dispatch rule is used during the non-surge phases and CR is used during the surge impulse (week 53 to 57). We observe that the cycle time during the surge is only slightly improved compared to the curve where FIFO is used during the whole simulated period (labeled “FIFO”). If, however, the CR rule is used up until four weeks after the surge impulse (curve labeled “FIFO CR FIFO (2)”, the best cycle time performance can be achieved. During the non-surge phase cycle times are as low as if using pure FIFO, and during the surge phase the cycle times are the same or even better than for the case where CR is applied during the whole simulated period (curve labeled “CR”). This can be explained by the fact that it takes approximately four weeks until the last lot that was started into the fab during the surge phase has left the fab. During this time, it is advantageous to use the CR rule since it can better deal with the high load during the surge impulse.

In Figure 6 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 then reset to the original value.

A clear peak of the WIP level is 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. Compared to the surge product, it also takes longer until the WIP levels return to their original values.

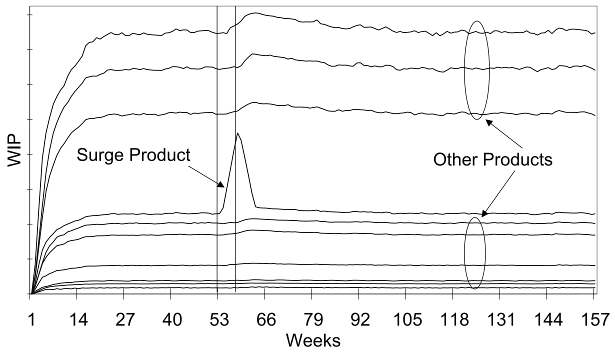

Figure 6: WIP Levels

In contrast to the WIP evolution, the cycle times of the different products, depicted in Figure 7, exhibit a parallel evolution. The relative increase in cycle time is about 5% to 6% for the different products.

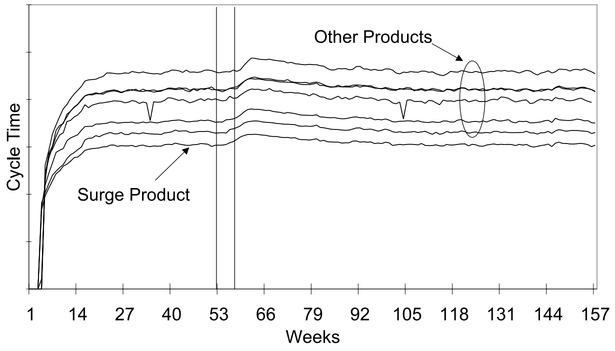

Figure 7: Cycle Times

Observing the evolution of the average number of finished wafers of the different products in Figure 8 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 leads to an overload situation at the bottleneck workcenter.

The most important conclusion drawn from Figure 8, 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 of increasing the start rate due to incoming orders.

4.2 Weekly Changing Product Mix

In reality, 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.

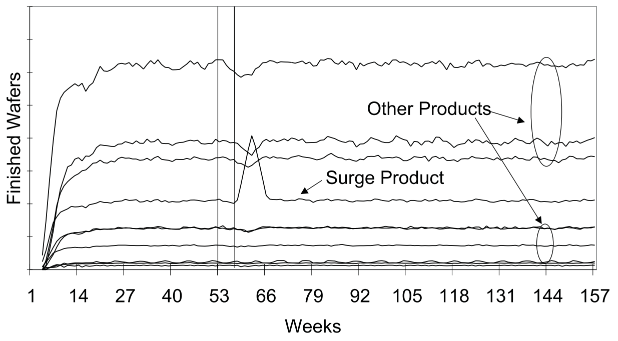

Figure 8: Wafers Out

In the first experiment, start rates are computed for each week independently from the start rates of the previous week and therefore are uncorrelated. To be more specific, the start rate $r_{i,j}$ for product i (where $i \in \{1, 2, \ldots, 10\}$) in week j (where $j \in \{1, 2, \ldots, 156\}$) is computed according to

$$r_{i,j} \quad = \quad r_{i,0} + f_1 \cdot U(-1, 1)\,,$$

where $r_{i,0}$ is the initial start rate of product i, f_1 is a factor that specifies the maximum deviation from the initial start rate, and $U(-1, 1)$ is a random variable sampled from an uniform distribution on the interval $(-1, 1)$.

In Figure 9, the cycle time curves of two sample products are displayed. We considered the four cases $f_1 = 0$ (constant product mix), $f_1 = 10\%$, $f_1 = 20\%$, and $f_1 = 30\%$.

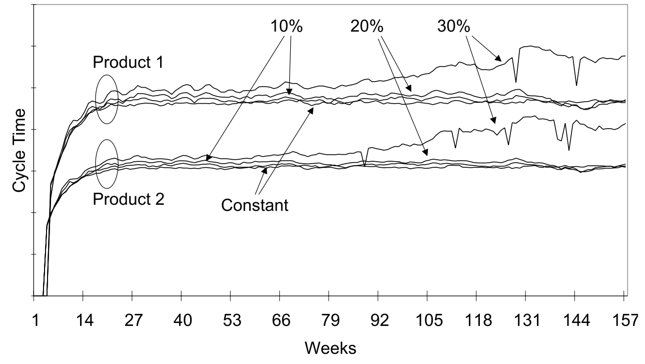

Figure 9: Constant vs. Changing Product Mix (Uniform)

The target start rates of each product, i.e., the three year average of the number of lots started, are almost identical in all cases. For $f_1 = 10\%$ and $f_1 = 20\%$, the cycle times do not differ much from the cycle times obtained in the case of constant start rates. However, allowing a maximum deviation of 30% in the start rates leads to a fab model that is no longer stable.

Table 1 contains performance characteristics for the four cases considered. Average cycle times are given as percentages of the cycle time for the case of constant start rates. The column "Tardy Lots" displays the percentage of lots that leave the fab after their due date. The final column contains the average lateness of the tardy lots.

Table 1: Weekly Changing Product Mix (Uncorrelated)

Max. Deviation	Average Cycle Time	Tardy Lots	Avg. Tardy Time of Tardy Lots (Days)
0%		0.70%	0.35
10%	102%	1.69%	0.55
20%	104%	20.17%	0.49
30%	112%	71.20%	1.13

As an alternative way of introducing weekly changes of the start rates, we consider an autoregressive (AR) process. In this model, the start rate $r_{i,j}$ for product i in week j is computed according to

$$r_{i,j} = r_{i,0} + f_1 \cdot (r_{i,j-1} - r_{i,0}) + f_2 \cdot r_{i,0} \cdot N(0,1)\,,$$

where $f_1 \in [0,1]$ determines how the start rate of the previous week influences the start rate of week j and f_2 determines how much variation is caused by the random variable $N(0,1)$, which is sampled from a standard normal distribution.

Figure 10 displays the average cycle times of a sample product for three parameter settings: f_1 was set to 0.7, f_2 was varied between 0.01, 0.05, and 0.1. As in the case of uncorrelated changes in the start rates, it is obvious that the statistical properties of the lot release scheme have a significant impact on fab performance.

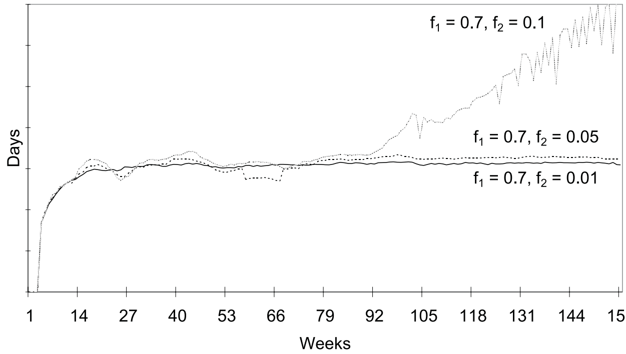

Figure 10: Changing Product Mix (Autoregressive)

4.3 Bulk Starts

In a third study we consider two different ways of releasing wafers into the fab: The first way is to distribute the lot releases evenly over each week, the second way is to release all the lots for a week in a single bulk at the beginning of the week.

For this study we keep the number of lots of each product released during one week constant for the simulation period of three years. Two different scenarios were simulated. In the first scenario, the start rates are chosen such that the bottleneck achieves an average load of 86%. Figure 11 shows the resulting weekly averages of the cycle times of a sample product.

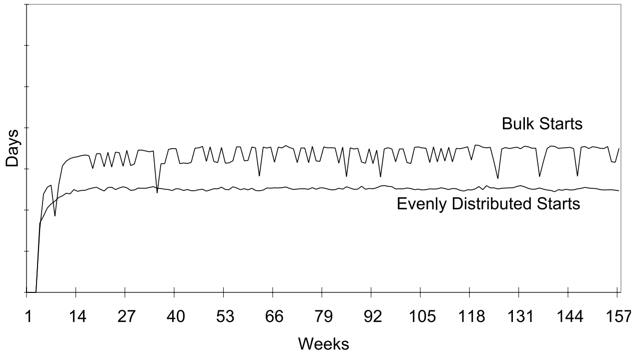

Figure 11: Bulk Starts vs. Evenly Distributed Starts (86% Utilization)

The cycle times of the sample product are considerably higher for the bulk start case. Also the variance of the cycle times is significantly higher, since the fab never can enter a kind of "steady state" due to the weekly bulk starts. Similar results can be shown for the nine other products.

Figure 12 displays the results for a scenario where the average bottleneck load is 93%. Three different cases are considered in this scenario: Evenly distributed lot starts under the WorkAPD dispatch rule, bulk starts under the WorkAPD rule, and bulk starts under the FIFO rule.

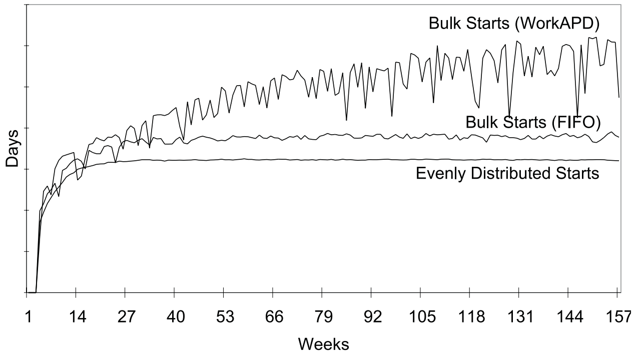

Figure 12: Bulk Starts vs. Evenly Distributed Starts (93% Utilization)

It is obvious that under a higher load the fab is not able to deal with the weekly bulk starts using the WorkAPD rule, since cycle times keep increasing during the three years. On the other hand, under the FIFO rule, the cycle times are not only lower than under the WorkAPD in the bulk start case, but also less variant.

In Table 2, some of the performance characteristics are summarized for the different scenarios. In the table, only the results for the experiments using the WorkAPD rule are displayed. Average cycle times are given for the bulk start cases as percentages of the cycle time for the corresponding case of evenly distributed lot starts.

Table 2: Bulk Starts vs. Evenly Distributed Starts

Bottleneck Load	Lot Starts	Average Cycle Time	Tardy Lots	Avg. Tardy Time of Lots (Days)
86%	Evenly		0.02%	0.22
	Bulk	140%	99.77%	3.86
93%	Evenly		21.08%	0.54
	Bulk	162%	100%	24.59

5 CONCLUSION

This paper presented results of a series of studies performed to analyze the impact that changes in product mix have on fab performance. We considered three types of changes:

- a surge scenario where the start rate of a single product is increased for a certain period of time,
- a scenario where the start rates for each product are changed at the beginning of each week, and
- a scenario where bulk lot starts at the beginning of each weak were used.

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. We also showed that it might be advantageous to change dispatch rules if the load situation in the fab changes.

Considering the case of weekly changing product mix showed that the statistical properties of the lot release scheme have significant impact on fab performance. Hence, production planners should not only take into account the capacity of the fab but also the variance of the lot release process. Comparing two different lot release schemes revealed that bulk start can drastically increase lot cycle times, especially if fab load is high.

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 J. Morrison. 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.

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

Dümmler, M., and Rose, O. 2000. Analysis of the Short Term Impact of Changes in Product Mix. *International Conference on Modeling and Analysis of Semiconductor Manufacturing (MASM 2000)*, 133-138.

AUTHOR BIOGRAPHY

MATHIAS A. DÜMMLER is a research fellow at the Department of Distributed Systems, Institute of Computer Science at the University of Würzburg, Germany, where he received his Master’s degree in 1997. He is interested in the modeling, analysis and control of manufacturing systems, especially in the area of semiconductor manufacturing. His main fields of research are batch service systems and cluster tools. He is a member of SCS, INFORMS, and GOR (German Operations Research Society). His email address is `<duemmler@informatik.uni-wuerzburg.de>` and his web address is `<www-info3.informatik.uni-wuerzburg.de/staff/duemmler>`.