

Measurable Improvements in Cycle-Time-Constrained Capacity

John W. Fowler, Ph.D.
Arizona State University
Tempe, Arizona, USA

Steven Brown
Siemens AG
Munich, Germany

Hermann Gold, Ph.D.
Siemens AG
Regensburg, Germany

Alexander Schoemig, Ph.D.
Siemens AG
Regensburg, Germany

Abstract – This study uses simulation-based analysis to evaluate the operating practices of a high-volume, multiple-product semiconductor fab, with the goal of finding potential areas for productivity improvement that will yield a quantifiable increase in fab capacity. The parameters setup, batching, tool/operator dedication, lot release, and dispatch rule were studied. The analysis revealed that some of the current operating practices of the factory were beneficial while changing some other practices would increase “cycle-time-constrained capacity” by up to 12%. A significant opportunity for potential improvement for this factory lies in implementing a strict setup avoidance policy. The first implementation in the actual fab is a relaxation of the photolithography equipment dedication that has helped the factory achieve a 25% reduction in cycle time and inventory.

INTRODUCTION

This study examines the factory parameters of setup reduction strategy, batching policy, tool/operator dedication, lot release policy, and dispatch rule as they apply to the practices of an existing Siemens wafer fab in Germany. These factors were chosen because they were thought to be influential and because they could be changed in the actual factory with minimal expense. Our specific goal [1] is to find potential areas for productivity improvement that will yield a quantifiable increase in fab ‘cycle-time-constrained capacity’, a concept we explain in a later section of the paper. To ensure that the recommendations for achieving this goal will be valid in the future, we evaluated wafer-start rates for two different time periods. First, the start rate plan for the first quarter 1997 was analyzed. Then, we repeated the experiments for the start rate plan for the last quarter of fiscal year 2000.

This was a joint project of Siemens AG, Arizona State University, and the University of Wuerzburg [6] and it leveraged heavily on previous work conducted with Dr. Frank Chance of Chance Industrial Solutions and Ms. Jennifer Robinson of the University of Massachusetts at Amherst [2]. All analysis is done with the capacity analysis and discrete-event simulation tool Factory Explorer™ from Wright Williams and Kelly.

METHODOLOGY AND SIMULATION APPROACH

Cycle-Time-Constrained Capacity

The relationship between the start rate and the average cycle time of products is given by the *Characteristic Curve* of the factory. Figure 1 shows two characteristic curves for a factory; one representing the way the factory currently

performs and the other representing the way the factory might operate under some changed (improved) conditions. The average cycle time has been normalized by dividing by the raw processing time (this is commonly referred to as a *multiplier of theoretical cycle time* or “*X factor*”). Note that for each curve, when the start rate is low (to the left of the chart) the average cycle time is close to the raw processing time. As the start rate increases, the average cycle time increases due to contention for resources (machines and operators).

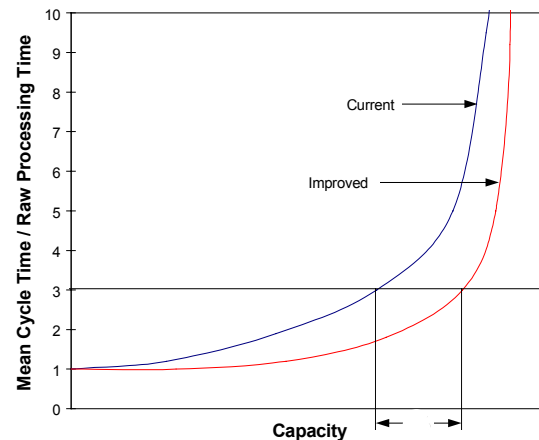


Figure 1: Cycle Time Characteristic Curves

In a previous study [4], *cycle-time-constrained capacity* has been defined as the maximum start rate (or throughput rate) sustainable for the factory for a given product mix, line yield, and equipment set, and a constraint on the average cycle time. To determine the cycle-time-constrained capacity for a given scenario, simply locate the desired average cycle time value and draw a horizontal line until it intersects the characteristic curve. At that point, draw a vertical line until it reaches the x-axis; the value found here is the cycle-time-constrained capacity. For example (see Figure 1), if an average cycle time that is three times the raw processing time is desired, we can determine the cycle-time-constrained capacity for both the current and the improved scenarios. The percent difference between these values measures productivity improvement. In this paper, the term *throughput* refers to the throughput obtained under a ‘multiplier of theoretical average cycle time’ constraint.

Experimental Plan

This analysis consisted of an initial set of screening runs to determine which of the five factors was most significant,

followed by a more detailed analysis of the critical factors. Each of the factors is described below.

The first factor addresses setups, which are modeled only at the implant steps. In the base case, no attempt is made to minimize the amount of setup; the next lot is chosen simply based on the default dispatch rule. To improve productivity, we consider a *setup avoidance* policy, a policy that will override the dispatching rule in order to avoid performing an extra setup on the machine. We refer to these policies as setup avoidance *off* and *on*, respectively.

The batching policy refers to the number of lots that must be present before a batch is started on a machine that can process multiple lots simultaneously, e.g., a furnace. We consider two batching policies: one that consists of only running full loads and one that runs a load as soon as at least one lot and a compatible furnace are available. These are called the *full* and *greedy* policies, respectively.

The existing operator grouping and tool dedication of the fab are reflected in the base case model. We refer to the base case as having *high tool dedication* because very small tool groups are dedicated to running only particular process steps. We also consider a *low tool dedication* scenario, where tool groups are combined, and each group processes more types of steps. For example, in the base case there are 14 stepper groups, whereas in the low dedication scenario, this is reduced to 7 stepper groups. The low dedication scenario also includes reduced dedication of several plasma etch tool groups. Differences in operator grouping associated with the tool groups are made accordingly.

We consider two lot release policies. The base case, referred to as the *Siemens* rule, releases lots into the system in groups of three at constant time intervals. Weekly start levels are predetermined by management on a quarterly basis. We also consider a policy called *CONWIP* that attempts to keep a constant amount of work-in-process (WIP) in the fab [7]. Under *CONWIP*, once the WIP reaches a pre-specified limit, a new lot is only released into the fab when a current lot finishes processing (or is scrapped).

We initially considered several dispatch rules with the intent of choosing the best performing rule for the final experiments. In our base case model, lots are processed according to the dispatching rule provided in the Workstream shop floor control system. The *Workstream* dispatching rule is based on the due dates of the lots. The due date of an individual lot is set to the release time plus the average 'planned' cycle time of the corresponding product. We also examine a *first in first out (FIFO)* rule and a *critical ratio* rule in the screening experiment. The *critical ratio* rule is similar to the Workstream rule except that it simply determines the critical ratio (time remaining until lot will be late divided by remaining processing time) and does not, as the Workstream rule does, explicitly prioritize the lots according to whether or not they are (or will be) late.

Each run of the model was for a time period corresponding to three years of operation. Statistical data was compiled after the initial transient phase of the system (roughly six months). Five replications of each scenario were made and averaged to account for statistical variation. Our screening experiments consisted of analyzing numerous scenarios. We then reduced the number of factors for the final experiments, as described in the results section. The final experiments were done for both the current and future scenarios. The future scenario was developed by taking the current model and changing the start rates to the start rates (and therefore product mix) that are planned for the last quarter of the fiscal year 2000. We then ran the capacity model to determine the number of tools and operators required to sustain this start rate without letting any equipment be utilized over 90%. In this way, the 'future fab' equipment was balanced against the anticipated wafer starts.

RESULTS AND RECOMMENDATIONS

Screening Experiments - Current Scenario

During our initial simulation runs, several things quickly became obvious. First, in most runs we noticed that the Siemens lot release policy performed much better than the *CONWIP* policy. This surprised us, as we thought that the average cycle time would decrease under *CONWIP* [7]. We concluded that *CONWIP* did not perform well in this case because lots were released one at a time into the factory rather than in groups of three lots. Releasing lots in groups of three tends to work well for this fab because the first processing step takes place on a batch machine with batch size of three lots. Therefore, in our final set of runs, we only considered the Siemens release policy. However, we think that a *CONWIP*-like policy that only releases groups of three lots might perform quite well. This is a possible area to explore in future studies.

Second, we noticed that a full batch policy performed at least as well as a greedy policy for most of the experimental design points. Since this is typically not the case for wafer fabs [3,4], we looked at the batch steps in the actual fab and noticed that most of the furnaces are loaded fairly heavily, which is where the full batch policy works best. We understand that this is what the fab currently does and, based on our experimentation, we think this is reasonable. Therefore, in our final set of runs, we only considered using a full batch at all batch steps. We think that it is possible that some additional improvement might be made by considering a full batch policy at the most heavily loaded furnaces, but something less than a full batch policy at the furnaces that are not loaded quite as heavily.

Our third preliminary observation was that the Workstream policy is generally considerably better, in terms of average cycle time, than a strict FIFO policy. We had expected the Workstream policy to perform better in terms of 'on-time delivery', but were not sure how it would perform in terms of average cycle time. We were pleased with this result since

this is the policy that the factory currently uses. We also found that a *critical ratio* policy performed well, particularly when the fab was very heavily loaded. Therefore, we included this policy in our final experiment.

Final Experiments - Current Scenario

Our preliminary experiments indicated that Siemens' settings for the lot release and batching policy factors were best left unchanged. Therefore, these factors were dropped from the final analysis. The preliminary experiments had also shown that using a setup avoidance policy would significantly improve cycle-time-constrained capacity. We also dropped this factor from our analysis, by applying the setup avoidance policy for all of our final runs. The final experiments involved investigating the two levels of the dedication strategy and dispatch rule factors for a model with setup avoidance. Because the current factory does not use a setup avoidance policy, we compared the final results to the original base model (without setup avoidance) to measure the potential improvement.

Table 1 shows the results of our final set of simulation runs for the current scenario. The first two columns of this table indicate the two factor settings that are compared. The next column indicates the percent improvement in cycle-time-constrained capacity over the base case (without setup avoidance). The final two columns show the 95th percentile of cycle time and the standard deviation of the cycle time (normalized by the raw process time).

Tool Dedication	Dispatching Rule	Percent Improvement compared to base case	95 th Percentile of Cycle Time/RPT	Std. Dev. of Cycle Time/RPT
High	WorkStream	6.72%	6.87	0.48
High	Critical Ratio	7.56%	5.44	0.10
Low	Workstream	11.43%	6.24	0.25
Low	Critical Ratio	12.61%	5.70	0.09

Table 1: Results for Current Scenario

The first row of Table 1 contains the results for a scenario that is similar to the base case except that in this scenario setups are strictly avoided. In this case, we see a 6.72% improvement in input rate over the base case, due to the setup avoidance. The next row indicates that there is some slight additional benefit in switching to a critical ratio dispatching policy. The third row shows what would happen if the amount of tool dedication was reduced.

Notice, in the fourth row, that there is again some slight additional improvement by switching to the critical ratio dispatching rule. While we have not found this improvement to be statistically significant, we believe that the critical ratio rule may perform considerably better than the Workstream rule when the factory is very heavily loaded.

We also note that, in general, the 95th percentile of the cycle time and the variance of the cycle times follow the trend of the cycle-time-constrained capacity. That is, the lower tool dedication and the critical ratio rule lead to lower 95th percentile and variance values. The two measures are a reasonable surrogate for on-time delivery. Thus making the recommended changes should not only lead to better throughput, but also to a tighter distribution of cycle times and therefore to better on-time delivery performance. Note in particular that the critical ratio rule leads to a dramatic reduction in cycle time standard deviation.

In reality, the fab does sometimes use setup avoidance; the actual policy implemented in the implant area is not as simple as 'on' or 'off'. Our numbers represent an upper bound on the expected improvement, while the difference between them (12.61% minus 6.72%, for example) represents the effect of the factors alone.

Figure 2 shows the characteristic curves for: 1) the base case; 2) the case that shows the significant advantage of setup avoidance; and 3) the final case that shows the combined effects of setup avoidance, low dedication of tools, and the use of the critical ratio dispatching rule.

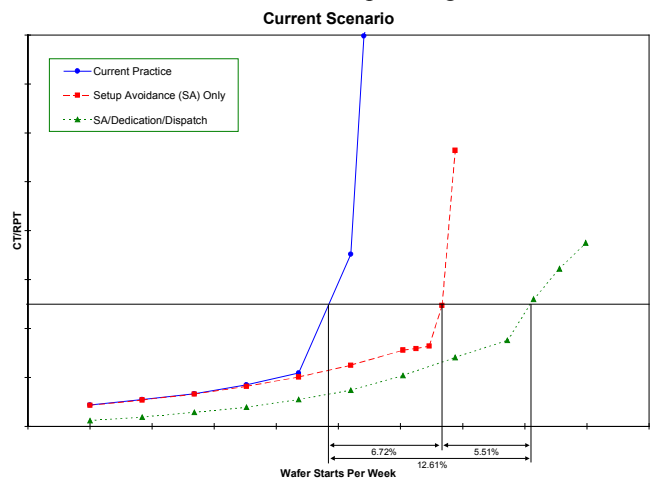


Figure 2: Characteristic Curves for Current Scenario

Final Experiments - Future Scenario

The final experiment (looking at dedication strategy and dispatch rule factors with setup avoidance) was repeated using the future scenario. Table 2 shows these results. Remember that when developing this scenario, we recalculated the equipment set for this new model (see previous discussion of the experimental plan). Therefore, we are assuming that the 'future fab' will add the new equipment recommended by the model and that this revised equipment set will sustain the new wafer start plan. In a like manner, operators were also added to support the additional equipment.

The results for this scenario are very similar to those for the current scenario, emphasizing even more the need for setup

avoidance; notice the 12.44% improvement over the base case just from using setup avoidance. Even considering an optimized tool set and a strict setup avoidance policy, the improved settings for the other two factors net an almost four percent improvement in throughput over the case with their original settings. This gives evidence that the recommendations will still have a positive impact on performance over the next few years.

Tool Dedication	Dispatching Rules	Percent Improvement compared to base case	95 th Percentile of Cycle Time/RPT	Std. Dev. of Cycle Time/RPT
High	WorkStream	12.44%	6.30	0.25
High	Critical Ratio	12.77%	5.86	0.09
Low	WorkStream	16.54%	6.54	0.09
Low	Critical Ratio	16.65%	6.30	0.08

Table 2: Results for Future Scenario

CONCLUDING REMARKS

The findings regarding tool dedication are of particular importance. This study, in agreement with previous analysis [2,6], shows that decreasing the level of tool dedication might significantly benefit this specific Siemens factory. In response to this recommendation, factory management pursued a change in operating methodology in the photolithography area that included relaxation of the stepper dedication policies. Within a short time, the factory realized a 25% reduction in cycle time (and, correspondingly, in inventory) without a reduction in throughput. The variability of the cycle times also decreased.

Certainly, we believed the tool dedication changes were a significant contributor to this improvement. Therefore, the modeling team ran additional simulations to compare the base case against the 'new' fab practice of relaxed stepper dedication rules (as opposed to the earlier experiments, where plasma etch tools also had relaxed dedication). Replications of three-year simulations showed a 29% reduction in the model's cycle time, which corresponds to a 3.7% increase in the model's capacity when cycle time is held constant. The average 95th-percentile of the cycle time distribution dropped by 21%, indicating that the model improvements not only come from a reduction in the *mean cycle time* but also from a reduction in the *variability* of the cycle time. Factory management is currently investigating the additional impact of relaxing the tool dedication rules in the plasma etch area.

This paper has presented a practical application of modeling and simulation to analyze the operating methods of a Siemens wafer fab. Simply put, we built a model of our existing factory, explored several options in production methodology, implemented a recommendation from the modeling analysis, and it made a difference. This case study

shows the benefit in applying performance analysis techniques to increase our understanding of the factory. By building and using a valid simulation model, improvements can be found and implemented into the real factory where they will do the most good.

ACKNOWLEDGMENTS

Special thanks to Ms. Jennifer Robinson, University of Massachusetts at Amherst, for her significant technical and editorial contributions to this paper. The authors also gratefully acknowledge the work of Dr. Frank Chance (Chance Industrial Solutions), Joerg Domaschke and Juergen Potoradi (Siemens AG), Shwu-Min Horng and Rueben Aguilar (Arizona State University), Manfred Mittler (IBM), and Oliver Rose (University of Wuerzburg).

REFERENCES

- [1] S. Brown, F. Chance, J.W. Fowler, and J. Robinson, 1997. 'A Centralized Approach to Factory Simulation', Future Fab International, June.
- [2] F. Chance, 1996. 'Factory Explorer™ Implementation Project Final Report', Siemens internal project specification and report.
- [3] J.W. Fowler, G.L. Hogg, and D.T. Phillips, 1992. 'Control of Multiproduct Bulk Server Diffusion/Oxidation Processes', IEEE Transactions on Semiconductor Manufacturing, Vol. 24, No. 4, pp.84-96.
- [4] J.W. Fowler and J.K. Robinson, 1995. "Measurement and Improvement of Manufacturing Capacity (MIMAC) Designed Experiment Report", SEMATECH Technology Transfer #95062860A-TR.
- [5] J.W. Fowler and J.K. Robinson, 1995. "Measurement and Improvement of Manufacturing Capacity (MIMAC) Project Final Report", SEMATECH Technology Transfer #95062861A-TR.
- [6] J.W. Fowler and P. Tran-Gia, 1996. 'Regensburg Productivity Potential Case Study Final Report', Siemens internal project specification and report.
- [7] M.L. Spearman, D.L. Woodruff, and W.J. Hopp, 1989. 'CONWIP: A Pull Alternative to Kanban', International Journal of Production Research, 28(5), pp.879-894.