ABSTRACT

Using discrete-event simulation models, a study was conducted to evaluate the current production practices of a high-volume semiconductor back-end operation. The overall goal was to find potential areas for productivity improvement that would collectively yield a 60% reduction in manufacturing cycle time. This paper presents the simulation methodology and findings pertaining to analysis of the Assembly, Burn-In, and Test operations. Many of the recommendations identified can be implemented at no additional cost to the factory. The most significant opportunities for improvement are in the Test area, the system constraint. Additionally, the model is extremely sensitive to changes in operator staffing levels, an accurate reflection of many back-end operations. The model shows that the cumulative impact of these recommendations is a 41% reduction in average cycle time, a significant contribution to the overall goal.

1 INTRODUCTION

The application of modeling and simulation for factory performance analysis is in the beginning stages in the semiconductor industry, relative to device and process modeling [Moore, 1997]. However, the National Technology Roadmap now identifies modeling and simulation as critical needs in the area of factory integration [Semiconductor Industry Association, 1997]. The Next-Generation Manufacturing Project Team goes even further in their discussion of pervasive modeling and simulation, predicting that all future production decisions will be made on the basis of modeling and simulation methods [The Agility Forum, 1997].

Modeling and simulation are becoming particularly critical for back-end operations. As yields and efficiencies from wafer fabs continue to increase, more attention is directed towards the ability of semiconductor back-end factories to handle the load with minimum capital expenditures. Throughput, utilization, and cycle time continue to be emphasized as key performance parameters for existing operations and for the complex planning of new facilities. Because static models cannot adequately handle this level of analysis, managers are turning more to simulation.

Figure 1 shows a typical semiconductor production flow. Siemens’ Semiconductor Division has recently used a simulation approach to address a classic capacity issue. The Dresden, Germany wafer fab has experienced a quicker-than-expected acceleration up the learning curve, resulting in higher yields than originally planned. This increase in saleable product is, of course, considered to be a good problem. The difficulty, however, is that the back-end equipment set now has a much higher production demand than planned for and has exceeded the original loading plan.

Figure 1: A Simplified Semiconductor Production Flow

The result is a much larger work-in-process inventory than anticipated and, consequently, unacceptable cycle times in the Assembly and Final Test operations. Management’s goal is to process the higher volume, within the original cycle time plan, without additional capital expenditures. To assist this effort, members from Siemens’ centralized Factory Modeling and Simulation Team [Brown, et. al., 1997] constructed a detailed model of the factory and completed a performance analysis of the production operations.
The project team’s goal was to identify changes in production operations that could collectively reduce back-end cycle time by 60% while maintaining the current capacity loading levels (see Figure 2). This was accomplished by applying simulation models and analyses.

![Figure 2: Goal of the Simulation Study](image)

An initial interview of the production management team revealed that they had established specific process enhancement goals designed to increase the throughput of the factory. Inputting this new data into an initial factory model showed that collectively these planned improvements would allow the factory to reach the new cycle time goal (see Figure 3).

![Figure 3: Changes in Cycle Time and Inventory with Planned Process Improvements](image)

Believing that all these process improvement goals might not be met, the simulation team decided to use this model to identify potential cycle time reductions through production management improvements. For this study, it was assumed that none of the planned process improvements would be achieved. The base model used current operating data for input.

Attempting such a dramatic cycle time reduction was a formidable assignment and required reviewing many different aspects of factory operations. Results of the study are expressed as a percentage decrease in the cycle time of the Assembly-and-Test combined factory. These recommendations show a cumulative 41% potential reduction in factory cycle time. The recommendations involve the impact of rework, test handler dedication, operator staffing, material batching, lot release, and transport lot size [Brown, et. al., 1998. Domaschke, et. al., 1998].

## 2 SIMULATION METHODOLOGY

### 2.1 Project Approach

The project began with the problem statement: How can the Dresden back-end factory reach its cycle time goal? With this in mind, the simulation team worked directly with the production department to determine the project deliverables, schedule, and milestones. For this project, the team used a project management approach established by Chance, Robinson, and Fowler [1996]. This methodology relies very heavily on production management involvement throughout all phases of the simulation project and embraces the philosophy of building the simplest model needed to answer the question.

### 2.2 Software

This project used the performance analysis software Factory Explorer, from Wright Williams and Kelly [Chance, 1996a], which proved to be a very effective tool for modeling back-end operations. Within thirty days a top-down model was constructed with available data. From this point, more detail was added to the model as needed to conduct an effective first-pass analysis of the existing factory. Working with key members of the production staff, this analysis and a subsequent, more detailed model were validated against actual factory output data. This process required approximately another two months. Additional input data, and more detailed data, were added to the model only as needed to answer specific questions.

### 2.3 Data

As in most simulation studies, the majority of the effort for this project was expended in collecting and preparing input data to construct a valid model of the factory. The team found that, in general, the required modeling data was already being collected and analyzed by the factory. This contributed significantly to the speed in building the initial model. One exception was the manner in which downtime data was collected. Factories that base production decisions on traditional static planning models typically express this data as a percentage of the time the machine is down for maintenance. To build the simulation model required the team to review historical records and express this downtime in terms of mean-time-between-failures (MTBF) and mean-time-to-repair (MTTR). This was initially a considerable effort but was later used as the basis for redefining data collection and data preparation procedures for shop floor control.
A key part of any data-collection procedure is to verify the accuracy of existing data. A 100% effort is not usually required, and is certainly not practical. However, visits to the shop floor to double-check input parameters for critical operations and equipment are very important to quickly building a valid model of the factory. Real-time observations and short conversations with operators, engineers, and supervisors can significantly contribute to early successes.

A major concern at Siemens Semiconductor is the amount of time simulation members spend on data management, as compared to actual simulation analysis. As a result of this project and concurrent wafer fab modeling projects at the Dresden factory [Peikert, 1998. Peikert and Brown, 1998], a standard approach to the data management problem was established. The basic premise is that the analyst should spend time performing modeling and analysis functions for the factory. The data management requirements should be left to the experts in other departments, such as Information Technology (IT) or Computer Integrated Manufacturing (CIM). A second key premise is that all data used for decision-making models should come from the same data warehouse.

Therefore, the approach developed at Dresden (see Figure 4) assigned to the IT and CIM departments the responsibility for transferring the data from various databases into a pre-defined data warehouse. Automatic transfer of detailed input data (e.g., process flows and product recipes) to keep the models current is their responsibility. The users of the models have the responsibility of defining the data and the data structure needed for their models. The principle here is that the centralized data warehouse is used for all modeling – static and dynamic – from simple spreadsheets to capacity planning to discrete-event simulation analysis.

2.4 Model Validation

The simulation team defines a valid model quite simply – it is a model that is accepted by production management to produce credible results and that is used in the decision making process. Partnering with production management at all levels is key to this acceptance.

A standard first-pass validation task is to select key output parameters for comparison. For this study, historical records of factory cycle time, cycle time by tool group, equipment utilization, and average inventory were compared against model outputs. Mean values from the model were found to be within 10% of historical values. In addition to this, the team often uses a very informal common-sense approach for validation. This involves observing the shop floor operations and conducting informal interviews with production personnel. The goals of this effort are to see where the model accurately represents reality and to understand why it does not. Key parameters generally investigated are the top ten cycle-time contributors and the system capacity constraints (where one might expect to see inventory queues). In addition to increasing model accuracy, this approach to validation gives the simulation analyst an opportunity to build credibility for the model among production personnel.

A good example of building credibility comes from a previous Siemens wafer fab simulation project [Chance, 1996b. Fowler, et. al., 1997], which investigated potential capacity improvements under a cycle-time constraint [Fowler and Robinson, 1995a]. As a part of this model validation, all furnace-area supervisors were asked the same question: "Do you have enough operators to run your area?". The common answer was "Yes, but I could run the area considerably better if I had just one more operator on my team". Analysis of the simulation output reports showed that the furnaces had an average waiting-on-operator time of 7%. This model-versus-reality comparison reinforced the simulation team's belief that their model was valid and also made the model output very credible to the furnace supervisors.

Once the Dresden back-end model was deemed by management to be a valid representation of the factory, the simulation team conducted a series of sensitivity analyses and presented the findings and recommendations. Results were typically expressed, as in Figure 2, via a graph of the operating curve for the factory. The operating curve shows the relationship between cycle time and capacity (or factory utilization). Improvements in cycle time create a shift in factory performance, which results in a new operating curve. The percent cycle time reduction for a given factory is the vertical difference between the two curves at the targeted capacity loading.

2.5 Partnering with Production

A key to the success of this (and any other) simulation project was the interactive exchange of ideas and information between the simulation experts and the production group. By working real-time with key production personnel, the simulation team could more easily produce meaningful findings and implementable recommendations. This investment of time and resources
by the production personnel allowed effective application of the modeling work.

Specific recommendations for cycle time reduction are discussed in the following section. Although no productivity improvement effort can be achieved without expending some resources, many of the changes described here can be implemented in the factory without adding capital equipment or incurring significant additional expense.

3 RESULTS

3.1 Test Procedure Change

One effective use of simulation is to take a planned improvement at one work station or operation and test the impact of this change on the performance of the overall factory. In this case, the management team was considering a process change at the Burn-In operation that would potentially eliminate the majority of the rework at the P2 Test operation (see Figure 5).

![Figure 5: Proposed Change to the Test Rework Process](image)

In effect, the change would move the rework loop away from the testers. Simulation analysis verified that this would indeed be a good strategy, resulting in a 12% decrease in overall back-end cycle time and a subsequent reduction in inventory (see Figure 6). Although this change significantly increases the process time at the burn-in ovens, the overall factory performance is improved because the testers are the bottleneck tools. Figure 6 clearly indicates the location of the factory constraint (high inventory level and significant increase in cumulative cycle time). This change effectively reduces the load at the system constraint, thus permitting the cycle time reduction, and can be implemented without additional capital expenditures. As a result of this analysis, production management implemented this process change for specified products.

3.2 Batching Policy at Burn-In

The Burn-In area currently operates under a full-load batching policy (an oven is left idle until enough inventory is available for a full load). Modeling the ovens with a greedy batching policy yields a 9% decrease in average cycle time when operating at the same capacity loading. A greedy policy requires each available oven to be immediately loaded with available inventory. The model shows this benefit to exist until the ovens reach 96% utilization, at which point cycle time begins to degrade. This result is consistent with results reported in Glassey and Weng (1991) and Fowler and Robinson (1995b), among others.

It is important to note that if these two recommendations were implemented together, the model shows that the Burn-In operation would become the bottleneck of the back-end factory. This highlights a very significant advantage of using simulation analysis to better understand factory operations. If incorporating changes to the production floor will cause the system constraint to shift, it is critical to know this in advance so that necessary adjustments can be made. Use of discrete-event simulation tools gives the manager this depth of understanding. In this case, knowing that the bottleneck could shift to Burn-In suggests a starting point to look for additional cycle time reductions in future studies.

3.3 Test Handler Dedication

Looking at the present bottleneck equipment group, the testers, current manufacturing strategy involves a high level of dedication due to production-imposed restrictions and to test handler restrictions. The model indicates that elimination of tool dedication at the factory bottleneck operation could yield a potential 11% reduction in cycle time (this result is heavily dependent upon the product mix). The cost to upgrade a test handler is equal to 8% of the tester’s original capital cost, much less than the cost of purchasing additional equipment. A similar result concerning the benefit of reducing equipment dedication is described in Fowler et. al. [1997].
3.4 Operator Staffing

The project model is very sensitive to staffing levels. This is not only an accurate representation of the Dresden factory but, the project team believes, also correctly depicts the typical back-end operation. By adding one operator per shift, and assigning "split" responsibilities between the Burn-In and Test areas (in effect, creating a "floater" operator), the average cycle time can be decreased by 8%. The factory implemented this change not through hiring additional operators, but by reassigning some tasks to available technicians and thus freeing up the needed operator time. These results indicate that there is great potential in conducting a detailed analysis of operator levels and qualifications throughout the factory. O’Ferrell [1995] also studied operator staffing and training, and found that cycle times in a wafer fab could be decreased by up to 50% through full cross-training of operators.

3.5 Lot Release into Assembly

Another very effective use of performance analysis software is in evaluating alternative scenarios of product mix and lot release strategies. Often such intricate changes to factory operations are cumbersome to initiate and impossible to track over the production cycle. They certainly can degrade factory performance. The way material is introduced into the factory (the mix of products, batching, and timing) determines downstream arrival distributions and greatly impacts the time a machine is idle due to material unavailability. In this analysis, the team simulated the current material release approach and compared it to various alternative combinations. The result was a recommendation for a new lot release strategy that has less variability and achieves a 12% reduction in modeled cycle time. The specific lot release strategy that is best is, of course, dependent upon the attributes of the given equipment set. Thus there is no universal application of an optimized rule. However, this analysis supports the general observation that actions to lower variability in a factory tend to reduce cycle times [Hopp and Spearman, 1996. Chen et. al., 1988]. Adjustments in lot release strategy have been made in the factory, focusing on the loading of downstream testers.

3.6 Transport Lot Size in Assembly

When the Assembly area was originally designed, it was expected that the production material would be transported from machine to machine in batches of 4,000 units. When inventory began to build due to increased production volumes, factory managers increased the transport lot size to 6,000 units. This meant that material was not moved to the next work station until all 6,000 units (12 magazines) were finished. Intuitively, the simulation team felt that this might be creating idle time as downstream equipment waited for the entire batch to be delivered. This idle time could reduce the available capacity of the downstream equipment. Further, having 12 magazines arrive at once to a downstream workstation that could only process one magazine at a time would lead to increased cycle times. A model of this actual scenario was compared to a second scenario in which the material was transported in batches of 500 pieces (one magazine) from the die bonders. The smaller batches were then recombined prior to leaving the Assembly area. The result was a 7% decrease in cycle time for the Assembly model and a 2% reduction for the complete factory model.

In effect, this change eliminated the routine idle time, and subsequent queuing, caused by waiting for material from machine to machine. The increased performance within the Assembly area delivered a smoother flow of material into the Test area and achieved the slight (2%) improvement in overall cycle time by decreasing the variability of the arrival rate to the system constraint. The impact is significant even though the system constraint for this factory is not in the Assembly area, but rather is in Final Test. This is because the factory sends material to additional sites for final test operations, and also sells some untested product to customers. Cycle time improvements at non-constraint work centers, as well as those at bottleneck machines, improve overall performance. This modeling recommendation has recently been implemented into the factory with expected results.

3.7 Plating Area Operator Scheduling

The available capacity of the galvanic, or electroplating, operation exceeds the capability of the overall factory by a large margin. Therefore, this operation has been run only on two shifts and remains idle at night. This, of course, greatly impacts the manner in which material arrives at the next production area. Again, the team was concerned that this variability might be creating idle time for critical downstream equipment. A scenario was modeled in which the same daily volume was spread linearly across three shifts, as compared to the current two-shift operation. This change allowed a smoother production flow through the factory. The model showed an 8% decrease in overall cycle time. Factory personnel on the night shift have been cross-trained so that this recommendation was implemented without hiring additional operators.

3.8 Overall Factory Cycle Time

Figure 7 shows the cumulative effect on the model if all these recommendations were implemented together. Since there are interaction effects present among the factors, the total impact is less than the sum of the individual effects.
In this case, the overall factory cycle time is decreased by 41% from the current levels.

This is a significant contribution to the established goal and, as previously mentioned, requires minimum additional expenditures. By treating the entire factory as a unified system, discrete-event simulation allowed the team to show the positive impact of effectively managing both constraint and non-constraint operations and equipment.

4 IMPLEMENTATION INTO THE FACTORY

There is little value to simulation for simulation’s sake. The goal of a manufacturing simulation project should be to have the modeling recommendations integrated into the decision-making process and implemented on the factory floor. The key to successful implementation is partnering with production.

For this project, the simulation analyst met regularly with key production and planning personnel to discuss progress and problems. As simulation milestones were met, project updates were given at routine weekly meetings. A key point of this standard project methodology is, where possible, to avoid appointments and special meetings. Discussions and presentations should be made during informal conversations and as an additional agenda item for an already-scheduled meeting. Information exchanges should be brief and to-the-point, recognizing that the most valuable asset your partner has is time.

The Dresden production management team has a routine weekly meeting to discuss production problem areas and long-term solutions. The simulation team was able to integrate their list of recommended improvements into this discussion of possible solutions and proactive initiatives. It is through this meeting that the five previously-discussed implementations were made.

5 CONCLUSION

Since this study began, the average cycle time for the Dresden back-end has decreased by 32%. Measuring the impact of implementing these specific recommendations is difficult because an aggressive fab production ramp and a change in product mix have created a factory of constant change over this time period. In general, however, the simulation team feels that the implementation of this work significantly contributed to this cycle-time decrease. The success of this project, and of the concurrent fab simulation projects [Peikert, 1998. Peikert and Brown, 1998], has led to a factory-wide acceptance of the benefits of simulation. The Dresden site has since formed its own simulation team for continued analysis. Guided by production management, the team consists of a simulation leader/coordinator and three full-time analysts investigating wafer fab, back-end, and transportation system topics. Additional positions are being considered for simulation support of the planning department and advanced manufacturing programs.

This case study shows the benefit of applying simulation modeling to performance analysis of a semiconductor back-end factory. Many aspects of the findings and recommendations made for this facility apply to complex manufacturing sites in general. For example:

- Unless batch machines are run at very high equipment utilizations, greedy loading policies usually lead to lower cycle times than full batch policies.
- Any change that increases the available capacity of the constraint workstation has a significant positive impact on the factory as a whole. This is consistent with the premise of the book *The Goal* [Goldratt, 1992].
- Equipment dedication, while often necessary for process reasons, tends to increase cycle times.
- Cross-training of operators, particularly those assigned to constraint workstations, can increase cycle-time-constrained capacity.
- Using smaller transport lot sizes reduces overall cycle times.
- Operating non-constraint equipment such that material flows smoothly to the constraint equipment is important.
- Lower variability in lot releases generally reduces cycle times.

Throughout this paper several important (in the authors’ view) points of simulation project management have been highlighted. A summary of these “lessons learned” is:

- Build the simplest model possible to answer the question(s) of interest.
- Validate the model with hard data from factory reports and with ‘shop floor intuition’.
- Partner with production, especially in the validation and implementation stages of the project. This is a key to project success.
- Respect the value of your partners’ time.
- Integrate your information-sharing into the factory’s existing meeting schedule.
Effective Implementation of Cycle Time Reduction Strategies for Semiconductor Back-End Manufacturing

- Do not try to sell simulation in general to factory management. Instead, show the benefits by addressing a specific concern/issue/problem.
- Do not allow the simulation analysts to get bogged down in database management – leave it to the experts.

More and more semiconductor managers are turning to discrete-event simulation to assist their decision-making for such complex factories. Recently there has been a significant increase in factory performance analysis papers at semiconductor manufacturing conferences. Such "simulation success stories" may be found in the proceedings of:

- International Symposium for Semiconductor Manufacturing (ISSM)
- Advanced Semiconductor Manufacturing Conference (ASMC)
- Semiconductor Manufacturing Operational Modeling and Simulation Symposium (SMOMS)
- Symposia in conjunction with annual regional SEMICON events, sponsored by the Semiconductor Equipment and Materials Institute (SEMI)
- SEMI Test, Assembly, and Packaging (TAP) Automation and Integration Conference
- International Electronics Manufacturing Technology (IEMT) Symposium

ACKNOWLEDGMENTS

The authors gratefully acknowledge the staff of the Siemens Microelectronics Center in Dresden, Germany for their assistance in data preparation, output analysis, and implementation of findings.

REFERENCES


AUTHOR BIOGRAPHIES

JOERG DOMASCHKE is a Simulation Engineer for Factory Modeling and Simulation, a part of the Manufacturing Science department of the Siemens Semiconductor Division. He has two years experience
conducting performance analyses for wafer fab and back-end factories. Graduating with distinction in 1996, he earned a Master of Science in Computer Science from the University of Wuerzburg. Queueing theory and simulation were his major subjects of study. His previous experience includes neuronal-network research for Mercedes Benz. Mr. Domaschke’s current simulation interests are in performance analysis of semiconductor factory systems.

STEVEN BROWN is the project manager for Factory Modeling and Simulation, a part of the Manufacturing Science department of Siemens' Semiconductor Division. He has twenty-nine years of management experience in various industries, including twenty years in semiconductor manufacturing. His formal education includes Bachelor of Science and Master of Science degrees in Business Administration. Mr. Brown is a member of IEEE. His simulation interests are in factory performance analysis.

JENNIFER ROBINSON is a senior analyst for C2MS Productivity Solutions. Prior to C2MS, she worked as an independent consultant in the semiconductor industry. Her clients in this industry have included SEMATECH, Digital Equipment Corporation, Seagate Technology, and Siemens AG. Ms. Robinson holds a B.S. (1989) in civil engineering from Duke University, an M.S. (1992) in Operations Research from the University of Texas at Austin, and a Ph.D. (1998) in Industrial Engineering from the University of Massachusetts at Amherst. Her research interests center on factory productivity measurement and improvement.

FRANZ LEIBL is the Director of Backend Operations for the Siemens Microelectronics Center in Dresden, Germany. He has been with Siemens for eleven years, with extensive experience in photolithography production management, having supported both pilot line and wafer fab operations. His recent work includes international advanced-technology programs.