IMPLEMENTING THE RESULTS OF A MANUFACTURING SIMULATION IN A SEMICONDUCTOR LINE

David J. Miller
IBM General Technology Division
Essex Junction, Vermont 05452

ABSTRACT

Semiconductor manufacturing is an extremely complex process. The requirements for ultraclean environments and intricate production capabilities result in capital intensive facilities and equipment. Wafer fabrication involves hundreds of individual tools performing multiple processes to produce an array of sophisticated end products. Worldwide competitive realities produce ongoing pressures to reduce the time and cost of both development and manufacturing activities. Simulation provides a practical and powerful method for analyzing and optimizing such a complex environment.

This paper presents an overview of a modeling and simulation effort designed to quantify turnaround time improvement opportunities in a leading-edge semiconductor development line. Product turnaround time is a key determinant of success in semiconductor manufacturing due to its contribution in critical areas including contamination, yield-learning, and process control.

The focus of the simulation project was the analysis of line loading levels and their impact on turnaround time. The results of the project led to policy changes and significantly improved turnaround time performance. The paper presents an overview of the line modeled, a discussion of the line characteristics incorporated in the model, and the simulation results. The changes implemented in the line are discussed, and the resultant improvements described.

1. INTRODUCTION

Product turnaround time (TAT) is defined as the clock or elapsed time from wafer release to completed wafer fabrication. It is also known as cycle time, mean elapsed time, or manufacturing lead time. It is a key determinant of success in the semiconductor industry due to its critical impact on contamination levels, process control capability, yield learning rates, and product costs.

Turnaround time is arguably more important in semiconductor fabrication than in any other industry. Semiconductor manufacturers must strictly control particulate contamination to achieve high device yields. Turnaround time directly contributes to the ultimate number of contamination defects on a wafer (Osburn, et al. 1988), as shown in Figure 1.

Process control is also critical to meet the intricate processing requirements associated with leading edge products. TAT directly affects ultimate process results. An example is the time between the onset of a problem and its ultimate detection and correction, where increased TAT magnifies the impact of a problem and delays the benefits of improvements. The slope of the yield learning curve is also a direct function of turnaround time. Figure 2 depicts the relationship between turnaround time and rate of yield learning.

In addition to yield impacts, turnaround time directly contributes to costs in such areas as space for storage, handling and tracking time, line control resources, and inventory carrying costs.
Success in semiconductor manufacturing requires more than quick turnaround times, however. The capital intensive nature of the facilities and equipment drives a requirement to maximize throughput to provide a competitive cost per wafer.

Line loading levels are a major determinant of both turnaround time and throughput performance. Throughput analysis techniques such as Groover (1980) and Buzacott (1971) generate curves such as Figure 3 that demonstrate that infinite loading provides maximum throughput.

![Figure 3. Throughput versus loading curve](image)

Queueing theory analysis (See Allen (1978) for example) produces the curve in Figure 4 where minimum loading produces minimum throughput times. The requirement to maximize throughput and minimize TAT thus leads to an inherent conflict in line loading decisions.

Actual turnaround time and throughput curves for a given machine depend on a large number of variables unique to that specific machine. The variables include such factors as arrival rates, service rates, rework rates, failure rates, and starvation and blockage opportunities. The analysis of hundreds of machines that make up a semiconductor fab becomes such a complex problem that simulation emerges as the most feasible and effective analysis technique available.

![Figure 4. TAT versus loading curve](image)

This paper presents the results of a simulation model developed for an existing semiconductor line. The focus of the project was to quantify the effect of line loading levels on turnaround times and throughputs within the constraints of the existing fabricator’s tool set, product mix, people resources, and control capabilities. The objective was to determine a load level that met line throughput requirements with minimum average turnaround time.

2. ENVIRONMENT

The simulation modeling project described in this paper took place in a leading edge semiconductor development line in IBM’s Essex Junction, Vermont facility in late 1987. The line modeled provides design verification, process development and demonstration of manufacturability for IBM’s leading edge memory semiconductors.

Turnaround time is especially critical in this environment where rates of feedback dictate rates of design and process advances. Throughput is also important to provide engineers and designers with sufficient hardware to verify development assumptions.

At the outset of the project, the line was moving product at an average of six times theoretical or raw processing time (RPT). RPT is a widely used measure of turnaround time, where RPT is the time it would take a predefined number of wafers to complete processing assuming no tool failures, no queuing, no rework, no waiting on operators, and no engineering holds. Thus RPT includes only the time required to process a lot, including setups, load and unload, and processing.

The primary product in the line required over 300 process steps in more than 100 different tools to fabricate a wafer. A percentage of the lots in the line were experiments that traveled through some subset of the process. The line was controlled by a combination of computer systems for lot tracking and reporting, and manual systems for lot release and priority assignments. Priority lots typically finished in less than six times raw process time, while lots without priority took longer.

3. METHODOLOGY

Since a line performance rate of 6X RPT was not adequate to meet new product development cycles, a project was initiated to analyze the line and define opportunities for improving turnaround times. There was no shortage of opinions on the subject, as various people maintained that the long cycle times were due to an engineering resource problem, a bottleneck at photo processing, a shortage of manufacturing operators, or excessive tool failures. Others believed that that there was too much work in process in the line, that there was not enough work in process in the line, or that the line ran as well as possible given the nature of development.
Recognizing the complexity of analyzing a semiconductor line, and desiring an objective analysis of existing and proposed performance, it was decided to build a model of the line and simulate the line under various conditions and assumptions. Burman et al. (1986) provides an excellent discussion of the complexity issue and the applicability of simulation in this environment.

Simulation was selected as the best available vehicle to produce quantitative results based on well-defined assumptions, independent of the various opinions, traditions, experiences, perceptions, intuitions, policies and previous positions that existed.

The project consisted of three major phases. The initial phase included an analysis of actual line performance and quantified key line parameters. The second phase was the construction of the line model and execution of line simulations, while the final phase was implementation of simulation results.

4. LINE ANALYSIS AND QUANTIFICATION

Line policies at the start of the project included some loosely defined loading targets which were not actively managed, and a quantity objective for wafer starts per day. A unique priority was assigned to every individual lot of wafers and manipulated by engineering and production control daily. The average line turnaround time performance was six times raw process times. A basic line tracking computer system was in place, along with a series of line performance reports on throughputs and turnaround times.

The first step was an analysis of the existing line performance. Turnaround time is generally measured as target or plan time versus actual or elapsed time. Plan time includes raw process time and planned time for factors such as transportation, queuing, rework, and resource availability. Actual time is the elapsed time required to process a predefined unit of work. The intent of this phase was to determine a realistic plan or target turnaround time for the line, and to quantify the major contributors to excessive turnaround times.

Analysis indicated that in a development line environment, there is an impact to TAT due to engineering holds and tool failures that is above what would be expected in a production environment. A realistic target for such a development line was determined to be 3X RPT. The breakout shown in Figure 5 includes 1X for RPT, 1X for manufacturing time to allow for factors such as rework, queuing, and operator availability, 0.5X for engineering to allow time for process development, and 0.5X for time lost due to tool failures.

The review of actual line performance data shown in Figure 5 indicated that the majority of the excess turnaround time was in the category labeled manufacturing time and not associated with excessive tool or engineering impacts.

![Figure 5: Plan versus actual TAT components](image)

The next step was the collection of line data on processes and equipment to ensure that raw process times and tool capacity plans were accurate. Unfortunately, but probably typically, much of the data was either unavailable or inaccurate. This led to a systematic review of the process and the tool set to define and verify process flow, process times, process tools, alternate process tools, tool availability, tool run sizes, lot sizes and so on. The result was a complete data base of process flow and tool information.

It was recognized early that maintenance of this data file would be critical for future use of both the data and the model. At the same it was apparent that the file contained all the necessary data to execute a process. Thus existing process descriptions were enhanced to include the data in well defined, formatted fields, improving the descriptions for line operator's use and providing the ability to directly create the central file from the most current process of record.

The benefits and applications of this phase went well beyond the subsequent modeling activity. The centralized line information file quickly became the basis for tool planning, layout planning, process assignment, and other line analyses in addition to the simulation uses.

5. MODEL DEVELOPMENT

Model development was driven by the project objectives. The model had to recognize the parameters of the actual process flows and installed tools. It then had to simulate the key characteristics of the technology development line (TDL) including tool failures, engineering holds, rework levels, process yields, variability in transport and material handling times, lot release schemes, priority processing, tool run sizes, lot sizes, lot batching rules, work in process storage areas, and alternate tool options.

The selection of simulation software was a critical step. A first pass using a generic line model was less than satisfactory as the model did not handle some very critical considerations including engineering holds, alternate tools, and...
and material handling. It also was extremely difficult to define and verify the model assumptions and logic. At that point, a review of available simulation software led to the conclusion that Systems Modeling Corporation’s SIMAN simulation language (Pegden, 1987) would meet the requirements of the project.

5.1 Model Structure

The SIMAN language divides a simulation model into an experimental frame and a model frame. The experimental frame contains the input data used by the model frame during the simulation. The model frame contains the unique logic used in the simulation. Thus the model frame stays the same and the appropriate inputs in the experimental frame are varied to generate various runs. See Pegden (1985) for a more complete review of the SIMAN structure.

5.2 Experimental Frame

The experimental frame was designed to provide the necessary line information in a straightforward manner. First, the SIMAN SEQUENCE element was used to describe process routings, with each tool assigned a unique station identification, and the process time, number of wafers to process, rework probability, rework loop and setup data defined as attributes for each process step in the routing.

The RESOURCE element was used to define the processing tools, work in process (WIP) storage racks, and operators. The SCHEDULE element provided the tool schedules, rack capacities and operator staffing levels by shifts. The PARAMETERS element described the statistical distributions used in the simulations. A PARAMETER was also created for each tool describing run size, primary storage area, alternate storage area, alternate tool, and planned availability.

The output reports containing cycle times, tool queue sizes, queue times, throughputs, operator’s utilization, tool availabilities, on-hold statistics, process yields, transport times and tool usage were defined in the experimental frame with the TALLIES, COUNTERS and DSTAT elements.

The vast majority of the experimental frame statements, including all of those described above, were generated directly from the process description files by a simple strip and format program. Other unique definitions including line control, priority, and loading levels were handled with the SIMAN INITIALIZE, RANKINGS, PARAMETERS or other experimental frame statements. Table 1 contains a summary of the experimental frame structure.

5.3 Model Frame

The model frame was divided into several major subsections to simplify the logic and provide necessary line performance simulation capability. The control of product flow in the model was based on the concept that the model had to determine where to go next for the lot (ENTITY) and had to determine what to do next for the tools and operators (RESOURCES) based on information defined in the experimental frame.

| INITIALIZE Element                      |
| Global Variable Initialization          |
| SEQUENCE Element                        |
| Process Routing Definitions             |
| Operation by operation with tool, run time, sample size, next tool in process, rework & setup information |
| TRANSPORT and DISTANCE Elements         |
| Transport and distance information      |
| RESOURCE Element                        |
| Define tools, operators, WIP storage units |
| SCHEDULES Element                       |
| Resource capacities and schedules for MTTF / MTTR, lot releases, operator schedules, etc. |
| RANKING Element                         |
| Define priority scheme in use at queues |
| PARAMETER Element                       |
| Statistical distributions for lot start, lot sizes, hold rates, process yields, etc. Also, used to store tool information such as run size, associated WIP storage units, alternate tool(s) |
| REPPLICATE Element                      |
| Simulation run definition: length, number of runs, etc. |
| TALLIES, COUNTERS and DSTAT Elements    |
| Define output reports and desired statistics |

The initial section used global variables to control releases including numbers of lots, product mixes, release frequency and lot sizes based on distributions and initialized values set in the experimental frame. It supported daily going rate or a fixed WIP release schemes. It also handled priority management logic, and initialized the appropriate parameters for collecting statistics for each lot.

Once a lot was released, it was routed to the processing section of the model using SIMAN’s STATION structure. The processing section executed the next operation in the routing as defined in the experimental frame SEQUENCE for every lot that entered the section. This section handled normal processing, including determining tool availability, selecting alternate tool options, calculating number of runs required, batching of lots, and seizing and releasing of required tools and operators.

This section also contained the logic for determining process yields, engineering holds, setup requirements and rework processing, again based on data from the experimental frame.

Finally, the processing section determined what to do with a lot once an operation was completed. The options included exiting the line if processing was completed or if no wafers were left, or going to the operation in the routing. Logic included determining where the next process occurred and deciding what type of transport was necessary. Lots were then sent to the next process step if a tool was available, or to a storage point if no tools were available to run the process.
The third major block of logic was added to manage the material handling and storage subsection. This group of stations used SIMAN DISTANCE and TRANSPORT mechanisms to move lots between areas, and to manage storage of lots on-hold or in queues for busy or down tools.

The last major section handled tool availability. The logic used the RESOURCE, SCHEDULE and PARAMETER information to determine the number of tools and planned availability. It then used statistical distributions to simulation tool failures throughout the line. Table 2 contains a summary of the model structure.

The model was written entirely in SIMAN language source statements and consists of several hundred lines of SIMAN source code. The experimental frame size depends on the number of routings, number of process steps per routing, number of unique tools, and number and types of output desired. The model runs on either the personal computer or the host versions of SIMAN, depending on the size of the experimental frame.

5.4 Model Verification and Simulation Runs

Initial model runs used actual data from the existing technology development line. The intent was to compare simulation results with actual line performance to verify the output of the model. The two largest initial variations were due to line operating procedures that were different in practice than on paper. Line operating procedures were improved as a result of these findings. Once the model was verified, a series of simulations was defined to analyze line performance under different lot release and line loading policies.

All runs in both the verification and analysis phases began with the line empty. The model ran for 100 days to load the line and then discarded the startup statistics. The model continued for a minimum of 150 additional days 50-day periods to generate statistics. A minimum of five runs was processed for each case using different random number seeds to generate a reasonable sample. In each case, runs were reviewed to assure that no new minimum or maximum points were generated as a basic indicator of line stability.

The major objective of the simulations was to define the optimum line loading level. A tool capacity plan existed that stated the tool set would support required throughput, but it did not consider turnaround time performance. The primary parameters varied in the simulations runs were: line loading levels, daily going rate versus fixed WIP release policies, and the impact of priority lots on overall turnaround times. The major output of the simulations were the points on the semiconductor line performance curves in Figure 6 defining the trade-off between load levels, turnaround times and product throughputs for the technology development line.

The simulation outputs and performance curves indicated that an 17% reduction in turnaround time was possible by reducing line loading with no other changes in tooling or staffing, and with no reduction in line throughput rates. Table 3 contains a summary of the

<table>
<thead>
<tr>
<th>Case</th>
<th>Max In Line</th>
<th>Throughput</th>
<th>Average Cycle Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Lots</td>
<td>#Prio</td>
<td>W/S/D</td>
</tr>
<tr>
<td>01</td>
<td>100%</td>
<td>20%</td>
<td>100%</td>
</tr>
<tr>
<td>02</td>
<td>100%</td>
<td>10%</td>
<td>98%</td>
</tr>
<tr>
<td>03</td>
<td>80%</td>
<td>20%</td>
<td>95%</td>
</tr>
<tr>
<td>04</td>
<td>80%</td>
<td>10%</td>
<td>102%</td>
</tr>
<tr>
<td>05</td>
<td>80%</td>
<td>5%</td>
<td>101%</td>
</tr>
<tr>
<td>06</td>
<td>70%</td>
<td>10%</td>
<td>100%</td>
</tr>
<tr>
<td>07</td>
<td>60%</td>
<td>10%</td>
<td>95%</td>
</tr>
<tr>
<td>08</td>
<td>50%</td>
<td>10%</td>
<td>79%</td>
</tr>
<tr>
<td>09</td>
<td>40%</td>
<td>10%</td>
<td>71%</td>
</tr>
</tbody>
</table>

† Optimum result: 100% of throughput at 83% of average TAT
results. Case 1 was the output of the model verification runs using actual technology development line data at the beginning of the project, and thus was the basis of comparison for all other runs.

The results further indicated that a fixed WIP release policy would provide better average TAT than daily going rate loading with no degradation to throughput rates over time. The output recommended smaller lot sizes to take better advantage of the growing proportion of single-wafer processing tools, and also suggested strongly that the existing priority scheme had little, if any, value, and actually might be impacting turnaround times by introducing additional variability into the line.

Other output included definition of pinch points, options for further trade-off of turnaround times versus throughput rates, operator utilization figures, queue size and wait time statistics, and tool utilization data.

6. IMPLEMENTATION OF RESULTS

Simulation was originally selected to analyze the line because of the fact that it could provide an objective analysis of existing and proposed policies independent of emotions, opinions, traditions, previous positions, experiences, intuitions and expectations. However implementation of results is very dependent on people with all of the above reactions. It is critical to develop credibility for the model before conclusions are presented. The approach used in this project was to verify the model results with several key people to develop that credibility before general policy changes were recommended.

The results of the simulation runs were summarized and presented in a series of reviews. The major recommendations were to reduce the number of lots in the line by 30%, to fix the line at a predefined number of lots with a fixed WIP lot release policy, to reduce the number of wafers in each lot, and to eliminate or drastically reduce use of priorities.

There were several concerns with the proposals that led to modifications before the results were implemented. For example, most people accepted the conclusion that the existing priority scheme was ineffectual and thus were willing to eliminate it. However, there was a need for priorities on critical design verification and process development lots. Thus the decision was made to limit the number of priority lots to a maximum of 10% of lots in the line, and to use only two categories: priority lots or normal lots.

Smaller lot sizes would improve turnaround time, but they would impact the flexibility for splits and multiple experiments in a development line environment. Thus, lot size reductions were not implemented.

There were concerns that reducing the number of lots would reduce engineering flexibility to run experiments, and would reduce manufacturing's ability to use people effectively. However, the simulation results demonstrated that the excessive lots sat in one of three queues: at bottleneck tools; at faulted tools; or awaiting engineering disposition. There was no predicted impact to either engineering or manufacturing productivity due to reduced WIP levels.

The decision was made to reduce the number of lots in the line by 20% initially. Once the simulated results were verified, line loading would be set to the 30% reduction recommended by the simulation output. At the same time a fixed WIP loading policy was implemented so that, once the line reached the target load level, no lots were released until a lot exited the line.

The results are presented in Figure 7. Actual line turnaround times fell an average of 16% over the first three months after the load was reduced 20%. The following quarter the number of lots in the line was cut to 30% below initial levels, and turnaround times fell an average of 25% from the initial 6X RPT rate to 4.5X RPT. Throughput rates actually improved slightly over the six-month period while the number of operators assigned to the line decreased.

![Graph showing product turnaround times](image)

Figure 7. Product turnaround time

Other benefits of the manufacturing modeling project included better balancing of workload across tools, improving procedures and data sources on the line, simplifying line management processes with improved reporting, and highlighting problem areas for further analysis.

7. MODEL APPLICABILITY TO OTHER FABRICATORS

While the model was developed for the technology development line, the model and frame layout are relatively independent of any specific fabricator. The model was built to simulate a semiconductor line, and several features were included that make it generic. First, it was purposely structured to support automatic loading of key data as described previously. This approach insured that data input was isolated from the model logic. Data has in fact been loaded from other lines in IBM using a strip and reformat program.
The model was constructed to support the major characteristics of a semiconductor line as described above, including process routings, product mixes, process yields, engineering holds, and tool failures. The SIMAN model and experimental frame structures provide an excellent vehicle for allowing input assumptions to be defined outside the model. This allows the simulation of various assumptions without changing the model. For example, rework, setup, engineering hold, or process yield loss logic can be turned off by setting the probabilities of occurrence to zero. Similarly, product mix assumptions, priorities and lot sizes are set by distributions. Adding or deleting tools and changing availability assumptions is also straightforward.

However, the model is generic only to the extent of the logic it contains. For instance, the model only recognizes lot sequencing policies supported by SIMAN. Other policies would require additional logic. Likewise, unique release policies beyond daily going rate or fixed WIP would require modifications to the model.

The model has been used to analyze other semiconductor lines, including the simulation of major new lines to predict line capabilities in terms of TAT and throughputs given a layout, tool set, and process descriptions.

8. APPLICABILITY OF RESULTS TO OTHER FABRICATORS

The loading versus turnaround time and throughput curves shown in Figure 6 demonstrate the importance of line loading. The technology development line realized a 25% improvement in TAT when line loading was reduced to numbers determined by simulation. The benefits were realized without any investment in additional tools, operators or complex management control systems, and without any impact to throughput rates.

The curves also define opportunities for additional turnaround time versus throughput trade-offs. Given the effect of TAT on final test yields, it is possible that more good chips could be produced at lower throughput rates in certain situations.

These results are important to every semiconductor manufacturing line. There exists a line loading level that provides a required throughput rate at a minimum turnaround time, given the constraints and policies of a specific line. Loading lines beyond that point to insure throughputs generally causes significant impacts to TAT with no compensating benefits.

Line loading is ultimately determined by lot release policy. The technology development line implemented a fixed WIP policy as a result of the simulations, fixing the number at the recommended load level. The results support the findings of Glassay and Resende (1988) that any reasonable closed-loop policy is better than rote release of product independent of line performance. Intuitively, a daily going rate policy tends to increase loading when lines perform poorly since less product is exiting the line, and tends to reduce loading when lines run well since more lots exit the line. Clearly this is contradictory to both throughput and turnaround time objectives.

Burman et al. (1986), Glassay and Resende (1988), Wein (1988), and Lozinski and Glassay (1988) have all presented excellent summaries of the complexity associated with the analysis of semiconductor manufacturing. They applied simulation to quantify the effect of various release and lot sequencing policies on the turnaround times of hypothetical lines, and reported results with release policies such as starvation avoidance and workload regulation at capacity bottlenecks that outperformed a closed-loop, fixed WIP policy. However, the technology development line does not have a control system capable of managing such policies at this time, and therefore they have not been attempted.

A general conclusion from their papers is that lot release policy has significantly more impact on average TAT that any subsequent lot sequencing policies. The results from the technology development line strongly support their conclusions. This is an important concept. It clearly suggests that fabricators should place primary emphasis on release policies beyond traditional daily going rate mechanisms, with lot sequencing and priority assignments a secondary consideration.

However, it should be noted that limited priorities do provide benefits in turnaround time for critical lots. For example, a policy with 10% of the lots designated as priority lots produced a 35% difference in turnaround time between priority and normal lots, while not affecting overall average TAT.

9. CONCLUSION

This paper has described a manufacturing simulation and the implementation of the results in a semiconductor line. The simulation project had a major impact on the management policies and actual performance of the technology development line. Line loading levels, release policies and priority mechanisms were all altered based on simulation results. Additionally, significant changes in line data collection, process description layouts, accuracy and availability of line information, and line reporting resulted from the modeling effort.

The simulation project provided tremendous insight into general semiconductor line performance considerations and significant information on the technology development line itself. It identified several other areas key to semiconductor line performance for further analysis and simulation. Overall, the project demonstrated the ability of simulation to analyze complex systems and demonstrated its potential to provide practical, meaningful results.
REFERENCES


AUTHOR BIOGRAPHY

DAVID MILLER is an advisory manufacturing systems engineer in IBM's General Technology Division. He received a B.S. in Mathematics and a B.S. in Psychology from Georgetown University in 1974, an M.S. in Computer Science from the University of Vermont in 1977, and an M.S. in Manufacturing Systems Engineering from Lehigh University in 1986. He joined IBM in 1974 and has held technology and management positions involved with advanced manufacturing systems at IBM facilities in Manassas, Virginia; Sindelfingen, West Germany; and Essex Junction, Vermont. His primary interests are in computer integrated manufacturing, factory automation, and production line modeling and simulation. He is a member of the Computer and Automated Systems Association of the Society of Manufacturing Engineers.

David J. Miller
IBM General Technology Division
Essex Junction, Vermont 05452, U.S.A.
(802) 769-8324