Proceedings of the 1999 Winter Simulation Conference
P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds.

ON THE CORRUPTING INFLUENCE OF VARIABILITY IN SEMICONDUCTOR MANUFACTURING

Alexander K. Schoemig

Infineon Technologies AG
Operational Excellence
P.O. Box 10 09 44
D-93009 Regensburg, GERMANY

ABSTRACT

This paper describes two simulation experiments using a model of a real medium sized multi-product semiconductor chip fabrication facility. The results presented clearly show the corrupting influence of variability, in this case caused by machine and tool unavailability. The immediate conclusion out of the results is that reducing the inherent variability of a manufacturing system improves the overall system performance. Hence, sampling shop-floor data should not only include first order statistics, but also measures that allow to monitor and model the variability of the machinery.

1 INTRODUCTION

Semiconductor manufacturing is among the most complex manufacturing processes as described by van Zant (1990). A semiconductor chip is a highly miniaturized, integrated electronic circuit consisting of thousands of components. Every semiconductor manufacturing process starts with raw *wafers*, a thin disc made of silicon or gallium arsenide. Depending on the diameter of the wafer, up to a few hundreds of identical chips can be made on each wafer, building up the electronic circuits layer by layer. Considering the scale of integration, the type of chip, customer specs, the whole manufacturing process may require up to 500 single processing steps.

Several performance measures are commonly used to describe and assess a semiconductor manufacturing facility. To highlight the most important of those we mention machine utilization, production yield, throughput, and last but not least cycle time. Cycle time is defined in this context as the time a lot of wafers needs to travel through the semiconductor wafer manufacturing process. In this study we do not consider wafer test, packaging/assembly, and final test.

Crucial factors of competitiveness in semiconductor manufacturing are the ability to rapidly incorporate advanced technologies in electronic products, ongoing improvement of manufacturing processes, and last but not least the capability of meeting due dates for an optimal customer satisfaction. In a situation where prices as well as the state of technology have settled at a certain level, the capability of meeting due dates along with the reduction of cycle time probably has become the most decisive factor to stand the fierce competition in the global market place. Consequently, operations managers are under a increasing pressure to ensure short and predictable cycle times.

2 BACKGROUND AND PROBLEM STATEMENT

2.1 The Operating Curve

In 1997 Infineon Technologies (then: Siemens AG‘s semiconductor division) started *the Productivity Offensive*. This project aimed at improving the capital efficiency of the 6-inch fabs located at Regensburg (Germany), Munich-Perlach (Germany), and Villach (Austria) by focussing on production logistics and. The making of semiconductor chips is a capital-intensive business. Consequently, endeavors requiring no or very little capital expenditure are undertaken to enhance production planning and control, in particular the reduction of cycle times to improve market response and on-time delivery. This consequence is especially a must for existing plants thriving on mature processes and products. Neglecting possibilities to enhance efficiency and productivity might push any existing plant out of business.

The *Operating Curve* Methodology (see Aurand and Miller (1997) and the references therein), also called *Characteristic Curve* as defined during the MIMAC project (cf. Fowler and Robinson (1995)), was introduced as the standard factory productivity measurement tool and a key performance indicator. Illustrative examples are given by Fowler et al. (1997) and Brown et. al. (1997).

As Figure 1 shows, the operating curve utilizes two metrics to benchmark and predict the performance of a manufacturing line: Mean cycle time and overall line

throughput. It illustrates the performance of the manufacturing line for the time period during data was sampled („*current*“) and predicts cycle time when the fab load (average amount of work released into the manufacturing line), fab capacity or variability due to improvements is changed („*improved*“). 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. The functional interdependence between cycle time and throughput is approximated by the Pollaczek-Khintchine formula (cf. Kleinrock 1975 p. 167ff.).

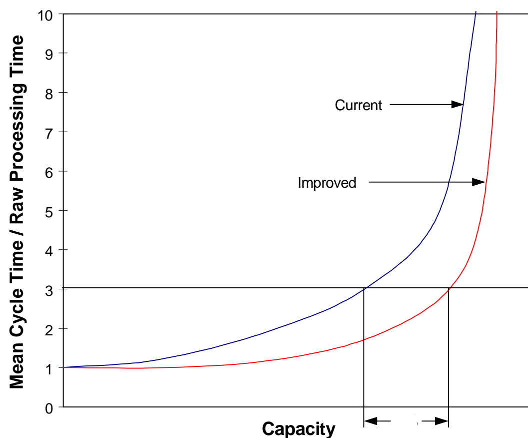

Figure 1: Examples of Operating Curves

Part of the reorganization of the operations management in the 6-inch fabs was the formation of crossfunctional work teams. These were before successfully introduced at the 8-inch Advanced Semiconductor Line (ACL) in Corbeil-Essones, France, which is jointly operated by Infineon Technologies and IBM (cf. Boebel and Ruolle 1996). Each team is burdened with the responsibility to maximize the efficiency of a certain manufacturing area. This includes problem localization, definition and execution of action plans, as well as long term problem monitoring and solving. Hence, all teams were trained in using special production data retrieval and visualizing software, applying the operating curve methodology, or, in other words, understanding the *Factory Physics philosophy* as formulated by Hopp and Spearman (1996). It was a surprising experience for some of the trainers, that most team members already had a good intuitive understanding of fundamental laws of production logistics like Little's Law or the non-linear nature of the operating curve, but lacked a sense for *the corrupting influence of variability* (Hopp and Spearman 1996 p. 282ff.) as a main performance detractor. Traditionally, the availability of a tool or machine along with the process speed has been regarded as the only parameters that determines its dynamic performance. (Atherton and Atherton 1995 p. 210ff.)

In advanced literature on queuing models, e.g., Takagi (1991), numerous examples and formulae are presented on the mathematical treatment of service systems with server vacation or breakdowns. However, almost all of these are cumbersome to use or even numerical intractable. Hopp and Spearman (1996 p. 266f.) provide a simple approximation for the additional variability introduced into a queuing system by server breakdowns:

$$c_e^2 = c_0^2 + (2(1+A)A\, m_r)/t_0$$

where A is the availability of the tool, m_r the mean time to repair, and t_0 and c_0^2 the mean and the coefficient of variation of the raw processing time.

This formula may be useful for illustrating the basic problem, but can not be applied when a sound understanding of the overall system performance is desired.

2.2 The Challenge of Operations Management in Semiconductor Manufacturing

A typical semiconductor chip manufacturing facility contains hundreds of various machines and tools such as masks used for lithography. Few machines are used for only one dedicated processing step. Most machines are designed to carry out several very similar processing steps during the whole processing sequence and for multiple products. Machines of the same type are usually grouped into *work centers* for several reasons: Reduction of setup time, redundancy in case of breakdowns, efficient utilization of operators, and having backup when maintenance work is done. Production control and operations management are tied to the flow of materials and the set of operations that transform raw material into the final products. There are several factors that make production planning and control in a semiconductor chip manufacturing facility particularly difficult. Hogg et. al. (1991) as well as Uzsoy et. al. (1992) summarize these factors thoroughly.

Given the complexity of the manufacturing process, carrying out scheduled maintenance as well as taking care of random machine breakdowns play a crucial role in semiconductor manufacturing. Despite of all efforts to tune and calibrate machines to an optimum performance, they are still subject to random failures. Obviously, downtimes are a severe problem, because the flow of material is disrupted and production capacity is lost. Unpredictable machine downtimes are believed to be the main source of uncertainty in the semiconductor manufacturing process.

There are several approaches to fight the effects caused by the randomness introduced into the manufacturing line:

- scheduling and sequencing of lots waiting for processing,
- dispatching rules and input regulation methods, and
- the control of the inventory.

Uzsoy et. al. (1994) describe the characteristics of various approaches to the shop-floor control problem in semiconductor manufacturing. The research on this topic is reviewed and classified, and the relative advantages and disadvantages of the solution techniques used are discussed. Mittler (1996) provides a broad investigation on modeling and analysis of variability factors and their impact on lot cycle times in semiconductor manufacturing. Chapter 5, in particular, focuses on equipment failure and repair and the Machine Interference Problem. However, despite all efforts to fight all negative effects of variability on cycle times, the efficiency of those methods seems to be very limited as Mittler et. al. (1995) show.

The aim of this spectrum of research is examining the concepts behind flow control heuristics and evaluating their effectiveness, overhead requirements and implementability and not the effectiveness of countermeasures on the machine level for reducing machine failures and consequently variability. Although the improvement of machine availability has always been a goal on the shop floor, little is known about the effect of *reducing the variability* caused by downtimes on the cycle time constrained capacity while the availability of machines might remain on the same level. In the following, two experiments are reported where the effect of a change in variability on the overall manufacturing line is observed.

3 RESEARCH METHODOLOGY

3.1 Simulation Model and Parameters

This investigation was conducted using the ***Factory Explorer***™ (FX) simulation tool, a package for capacity analysis of large manufacturing systems, with an emphasis on providing building blocks for modeling semiconductor manufacturing. FX combines an Excel-based interface with two performance analysis engines – one utilizing queuing formulae and one containing an discrete event simulator. Prior to September 1st, 1995 these FX engines were known as *Delphi*. This tool was used during the 1994 joint SEMATECH / JESSI project MIMAC (Measurement and Improvement of Manufacturing Capacity) see Fowler and Robinson (1995).

In 1996/97 a detailed model of the Regensburg multi-product semiconductor fabrication facility was built using FX. Features modeled include:

- 10 different process flows (4 memory, 6 logic products),
- operators,
- scrap and rework,
- dynamic dispatching (Workstream™ APD),
- lot transportation,
- sequence-dependent set-up times,
- recipe-dependent batching, and
- machine unavailability due to failures, preventive maintenance, and engineering.

This simulation model was used in the past two years for several studies on tool dedication, hot lots, and operator staffing levels to mention the most important ones. This model is agreed to be valid.

Each run of the simulation model for this study was for a time period corresponding to three years of fab operation for generating the operating curves and five years to sample cycle time distribution data. In any case, statistical data was sampled only after the initial transient phase of the system, what is roughly six months for a stable system.

FX utilizes the Schruben test to detect initial bias in simulation output. Briefly, this test forms a test statistic that is sensitive to changes in the batch means, the method used in FX to average output and generate confidence intervals. This test statistic converges in a statistical distribution of a known characteristic against which the empirical distribution of the actual output can be tested.

3.2 Experimental Design

Screening experiments consisted of analyzing numerous scenarios and parameter sets. For this presentation the number of factors is reduced and the major effects are highlighted.

The first experiment answers the question „what is the impact of changing variability, caused by machines, on the overall manufacturing line performance“, measured by the operating curve and cycle time distribution. FX provides a convenient run-time option that allows the user to multiply all machine and tool interruption time-to or units-to and time-offline mean values by a certain factor, while the percentage of time the tool or machine is unavailable due to a interruption (failure, PM, engineering, etc.) does not change. By this means the user affects only the frequency and severity of interruptions and not the theoretical maximum manu-facturing capacity of the tools and machines. Please note, that the effect modeled here is

reflected in Hopp and Spearman's formula. In this experiment the down time parameters were doubled (Exp1A) and halved (Exp 1B).

The second experiment is concerned with exploring the effect of the distribution of down events. In the base simulation model the usual assumption is made, that the random variables time-to-fail (if not based on the consumption of materials) and the time-to-repair are exponentially distributed. In this experiment, downtimes were changed to distributions as summarized in Table 1.

Please note here, that the properties of the statistical distribution of downtimes are not used in Hopp and Spearman's formula, and hence, can neither be used for calculations for educational nor for capacity considerations.

Table 1: Statistical Distributions of Tool and Machine Downtimes

	Base Case	**Experiment #2**
Downtimes (failure)	Exponential	Erlang-4
Engineering	Exponential	Triangular, +/- 10% of mean
Preventive Maintenance	Exponential	Triangular, +/- 10% of mean

4 RESULTS

4.1 Experiment #1

Figure 2 displays three operating curves: the base case, and two for the first experiment, where the experiment the down time parameters were halved (Exp1A) and doubled (Exp 1B). For low to medium system load there is obviously no or only little effect. When lot release into the system approaches the maximum capacity of the bottleneck machine group, cycle time increases in a non-linear fashion. This increase is the more distinct the higher the variability in the system is. Keep in mind that the availability of the tools and machines is in all three cases the same and hence the static capacity is the same! When we think in terms of *cycle time constrained* capacity, however, the experimental parameter under observation has an tremendous impact.

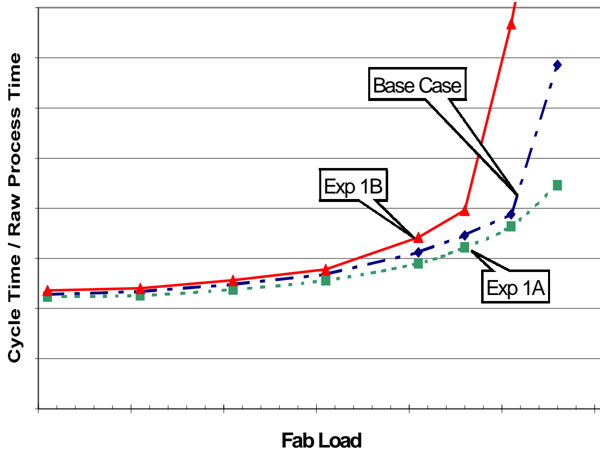

Figure 2: Operating Curves for the Base Case and Experiments '1A' and '1B'

In Experiment 1B the cycle times obviously spin out of control at a fab load where the fab safely operates in the base case. Analogously, the fab is able to bear a higher load, when the variability contribution of the tool is reduced by 50%.

Figures 3 and 4 give two examples out of the ten processes how variability impacts the cycle time distribution of lots. With increasing variability in the fab, not only the mean of the cycle time increases, also the distribution of cycle times spreads out, what is of course not desirable from an operations manager's point of view. This spreading effect is clearly visible for the data of the product depicted in Figure 3. Nevertheless, other products might not be affected to this contend as it can be seen in Figure 4.

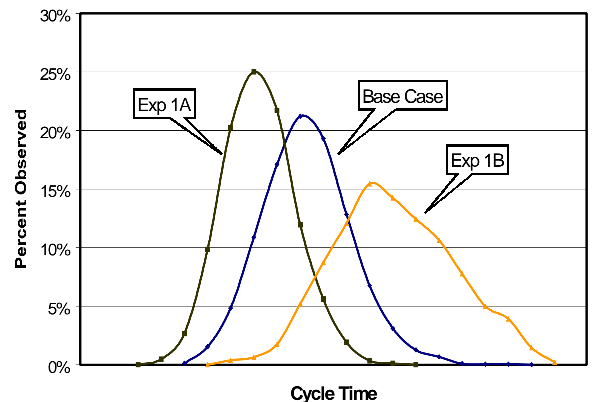

Figure 3: Spreading Effect of Cycle Times of Lots for a Particular Product

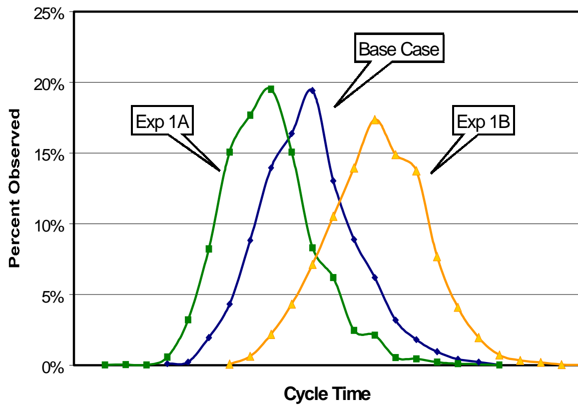

Figure 4: Second Example of the Spreading Effect

4.2 Experiment #2

The operating curves depicted in Figure 5 show no significant difference even when the system reaches a high load. Hence, we conclude that in this case the actual distribution of downtimes play only a minor - if not negligible - role in the performance of the fab. However, this conclusion should not be used as a justification of uncontrolled or even deliberate high variable downtimes. In any case, it must be concluded that we find here an open field for further research that goes beyond the very few cases considered in this study.

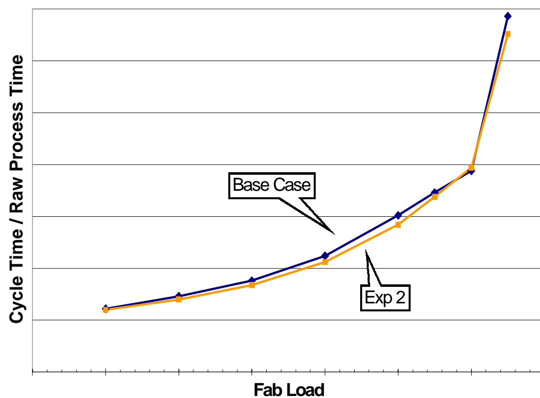

Figure 5: Experiment 2 Shows No Significant Impact of the Downtime Distribution

5 CONCLUSION

In this paper two simulation experiments were presented using a model of a real multi-product semiconductor fabrication facility. The results prove the corrupting influence of variability, caused by machine and tool unavailability, and also show the shortcomes of classical static capacity calculations. The main conclusion out of the results presented is that reducing the variability in the manufacturing system enables the ensurement of low and predictable cycle times. Hence, the precise sampling of shop-floor data, such as machine down times is a must. Theses statistics should not only include first order measures like means, but also statistics that allow to monitor the variability of the manufacturing system. Proper actions are advised if performance detractors are deducted and the effectiveness of these actions can be monitored and assessed using the same data visualization system.

REFERENCES

Atherton, L., and R. Atherton. 1995. *Wafer Fabrication: Factory Performance and Analysis*. Boston: Kluwer.

Aurand, S., and P. Miller. 1997. The Operating Curve: A Method to Measure and Benchmark Manufacturing Line Productivity. *IEEE/SEMI Advanced Semiconductor Manufacturing Conference,* 391-397.

Boebel, F.G., and O. Ruelle. 1996. Cycle time reduction programm at ACL. Proceedings of the *IEEE/SEMI Advanced Semiconductor Manufacturing Conference 1996,* 165-168.

Brown, S., F. Chance, J.W. Fowler, and J. Robinson. 1997. A Centralized Approach to Factory Simulation, *Future Fab International*.

Fowler, J.W., and J.K. Robinson. 1995. Measurement and Improvement of Manufacturing Capacity (MIMAC) Project Final Report. *SEMATECH Technology Transfer #95062861A-TR*. Austin, TX. Also published as*: Manufacturing Science and Technology for IC Production*, JESSI T30C / ESPRIT 8003, Theme 3.3, MST3-AI300-R-NI04-1.

Fowler, J.W., S. Brown, H. Gold, and A. Schoemig. 1997. Measurable Improvements in Cycle-Time-Constrained Capacity. *Proceedings of the Sixth International Symposium on Semiconductor Manu-facturing (ISSM)*, San Francisco, U.S.A.

Hogg, G., J.W. Fowler, and M. Ibrahim. 1991. Flow control in semiconductor manufacturing: A survey and projection of needs. *SEMATECH Technology Transfer #91110757A-GEN*, Austin, TX.

Hopp, W. J., and M. L. Spearman, 1996. *Factory Physics. Foundations of Manufacturing Manage-ment.* Chicago: Irwin.

Kleinrock, L. 1975. *Queueing Systems, Vol. 1: Theory*. New York: Wiley.

Mittler, M. 1996. The Variability of Cycle Times in Semiconductor Manufacturing. Ph.D. thesis, Bay. Julius-Maximilians-Universität Würzburg, Institut für Informatik. Published in: *Würzburger Beiträge zur Leistungsbewertung Verteilter Systeme,* Vol 1.

Mittler, M., A. Schoemig, and N. Gerlich. 1995. Reducing the Variance of Cycle Times in Semicon-ductor Manufacturing Systems. *Proceedings of the International Conference on Improving Manu-*

facturing Performance in a Distributed Enterprise: Advanced Systems and Tools, 89-98. Edinburgh, UK.

Takagi, H. 1991. *Queueing Analysis, Volume 1: Vacation and Priority Systems*. Amsterdam 1991.

Uzsoy, R., C. Lee, and L. Martin-Vega. 1992. A review of production planning and scheduling models in the semiconductor industry, Part I: System characteristics, performance evaluation and production planning. *IIE Transactions on Scheduling and Logistics* 24: 47-61.

Uzsoy, R., C. Lee, and L. Martin-Vega. 1994. A review of production planning and scheduling models in the semiconductor industry, Part II: Shop Floor Control. *IIE Transactions on Scheduling and Logistics* 26: 44-55.

van Zant, P. 1990. *Microchip Fabrication*. New York, 2nd ed.

AUTHOR BIOGRAPHY

ALEXANDER K. SCHOEMIG joined Infineon Technologies AG in 1997 (then: Siemens AG, Semiconductors) as a Simulation Engineer and Operations Analyst. He is currently responsible for the training of the simulation teams of the Infineon 6“ fab cluster and the integration of queuing modeling methods and simulation techniques in Infineon’s operations management system.

Dr. Schoemig received a Master degree in Computer Science in 1992 and a Ph.D. in Natural Sciences in 1997 from the University of Wuerzburg, Germany. From 1993 until 1997 he has been research fellow of the German Research Foundation (DFG), working in a project about stochastic modeling of manufacturing systems. His main research interests are modeling and performance evaluation of production systems and business processes using queueing theory, stochastic petri nets, and discrete event simulation.

Dr. Schoemig is member of INFORMS, GI (German Chapter of the ACM), and GOR (German Operations Research Society) .