Proceedings of the 2004 Winter Simulation Conference
R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

MODELING TOOL FAILURES IN SEMICONDUCTOR FAB SIMULATION

Oliver Rose

Institute of Computer Science
University of Würzburg
Würzburg, 97074, GERMANY.

ABSTRACT

In this research, we investigate how well Weibull, Gamma, and special bimodal distribution are suited as an alternative to the exponential distribution approach in the stochastic modeling of machine downtimes and times between failures. We also discuss the question whether sampling shop-floor data should not only include first order statistics, but also measures that allow to monitor and model the variability of the equipment and processes and even the correct distribution of the data.

1 INTRODUCTION

A typical semiconductor manufacturing facility contains up to 1000 various machines and tools. Besides the complexity of handling this vast amount of equipment, there are several other factors that make production planning and control in this environment particularly difficult. (Cf. (Schömig and Fowler 2000) and (Uzsoy et al. 1992) for a thorough summary of these factors and shop-floor control problems in semiconductor manufacturing.) Unpredictable machine downtimes are believed to be the main source of uncertainty in the semiconductor manufacturing process. Obviously, downtimes are a severe problem, since production capacity is lost and the flow of material is disrupted. The reliability of semiconductor manufacturing equipment is unusual from a number of standpoints. Despite of all efforts to tune and calibrate machines to an optimum performance, they are still subject to random failures. The failure of equipment or processes is often not a hard failure in the sense that something obviously breaks or goes wrong; but rather, a soft failure in which the equipment begins to produce out of the tolerance region. For this reason, the equipment usually completes a lot or batch prior to being taken out of service for repair which often involves more tuning, calibration, and test rather than component replacement. Since some wafer fabrication tools, such as ion implanters, may be down 30-40% of the time, the impact of periods of unavailability on production control as well as overall productivity is tremendous. Hence, appropriate modeling of equipment and process failures is a must to derive meaningful output performance measures. The SEMI E10 and E58 standards provide a framework for sampling machine-level data in the semiconductor industry.

Downtime is a period of time during which the equipment is not in a condition to perform its intended function. This period does not include any portion of time, where the equipment or the entire facility are not scheduled to perform fabrication. Generally, it is distinguished between *scheduled* and *unscheduled* downtimes.

A scheduled downtime occurs, when the equipment is not available to perform its intended function due to *planned* events such as preventive maintenance, production test, change of consumables, and machine setup for running a different process. All of these procedures are clearly separable and planned in their respective process. Also included are test run times for the required subsequent requalification and re-approval. Waiting times resulting from delays in the process are also included.

Unscheduled downtimes are periods of time during which the equipment is not in a condition to perform its intended function due to an unplanned event. Examples are: technical failures, unplanned measures to secure operation, unplanned shut down of supply infrastructure. These events interrupt equipment operation. In resolving these interruptions (Interrupts) they are distinguished as follows based on timing and personnel requirements:

An *assist* is an unplanned interruption that occurs during an equipment cycle if all three of the following conditions apply: (1) The interrupted equipment cycle is resumed through external intervention (e.g. by an operator), (2) there is no replacement of a part, other than specified consumables, (3) there is no further variation from specifications of equipment operation. An assist usually lasts not longer than 6 minutes. A failure, however, is any unplanned interruption or variance from the specifications of equipment operation other than assist.

2 MODELING EQUIPMENT DOWNTIMES

Obtaining the averages of uptimes and downtimes are sufficient when these time periods are assumed to be exponentially distributed. This is the prevalent assumption in reliability and simulation modeling when using simple models. Previous experiments (Schömig 1999) were concerned with exploring the effect of the distribution of down events. The results proved the corrupting influence of variability, that is caused by equipment unavailability, and also showed the shortcomings of classical static capacity calculations, the experiments concerning the type of the downtime distribution however, found no significant difference even when the system reaches a high load. Hence, it was concluded that in this case the actual distribution of downtimes play only a minor role in the performance of the fab.

Further experimentation in the context of an investigation concerning how simplifying assumptions in the stochastic modeling process for closed form queuing-type formulae affect the derived output performance measures as well as recent publications (Leemis 2001) gave reason to revisit the problem of finding an appropriate distribution for modeling the productive time between equipment failures and the time to repair.

In (Schömig and Rose 2003), we discussed the problems of finding the parameters for a Weibull failure model from real fab data. In addition, we provided some first results on the effects of different failure models on cycle times for a variety of fab models. We concluded that not only the mean and variance of the failure data plays an important role but also the shape of the distribution.

3 BIMODAL FAILURE MODEL

Because of several practitioners' comments on our aforementioned paper we investigated bimodal TTF (Time To Failure) and TTR (Time To Repair) distributions for our factory models. The principal motivation is that real factory tool failure measurement histograms are also bimodal in a lot of cases. This is due to the fact that there are a lot of short outages that happen frequently and a few long failures that occur rarely.

Due to the lack of real fab data, we developed a simple model for this type of machine failures where we can use given averages for the TTR and TTF and their coefficients of variation (CoV), where the CoV is defined as the standard deviation divided by the mean. The two CoVs are assumed to be the same for TTR and TTF.

For simplicity, we construct our bimodal distributions from two symmetric triangular distributions of type

$$\text{tria}(\min, \text{mode}, \max) = \text{tria}\left((1-\Delta)m, m, (1+\Delta)m\right),$$

where $0 < \Delta < 1$.

The mean of this triangular distribution is m, its variance is $\frac{\Delta^2 m^2}{6}$, and its CoV is $\frac{\Delta}{\sqrt{6}}$.

In the first part of model description we outline the computation of the individual average values for TTF and TTR of the two triangular distributions. In the second part, we determine the span/range of the triangular distributions to adapt the CoV of the bimodal distributions.

In the following, we introduce the notation of the parameters used:

Input parameters: Average TTR R and average TTF F, target CoV C.

Output parameters: Average TTRs r_1 and r_2, average TTFs f_1 and f_2, span of the triangular distributions Δ.

In the sequel, subscript 1 will be used for the short/frequent failures and 2 for the long/rare ones.

To compute 5 output parameters from 3 input parameters we need to introduce an additional assumption: Both the short/frequent and the long/rare failures lead to the same outage percentages $1-a_1$ and $1-a_2$ which are equal to half the outage percentage induced by the input parameters

$$1-a_1 = 1-a_2 = \frac{1}{2}(1-A),$$

$$\text{where } 1-a_1 = \frac{r_1}{r_1+f_1}, 1-a_1 = \frac{r_1}{r_1+f_1},$$

$$\text{and } 1-A = \frac{R}{R+F}$$

In addition, we need to relate one of output parameters to one of the input parameters. In our case we choose r_1 from the interval $\left]\frac{R}{2};R\right[$.

Smaller values of r_1 lead to larger CoVs and larger values to smaller CoVs of the resulting bimodal distribution.

After some algebra we end up with

$$f_1 = r_1 \frac{1+A}{1-A},$$

$$r_2 = \frac{R(r_1+f_1)}{r_1+f_1+(r_1-R)\frac{2}{1-A}}, \text{ and}$$

$$f_2 = r_2 \frac{1+A}{1-A}.$$

Based on these 4 average values, the probability α_1 that a short/frequent outage is computed results in

$$\alpha_1 = \frac{r_2+f_2}{r_1+f_1+r_2+f_2},$$

and analogously

$$\alpha_2 = \frac{r_1+f_1}{r_1+f_1+r_2+f_2}.$$

The variance of the bimodal distribution for the TTR generated from the two triangular distributions is derived as

$$\frac{6+\Delta^2}{6}\left(\alpha_1 r_1^2+\alpha_2 r_2^2\right)-R\ .$$

This leads to a minimum CoV of the combination of triangular distributions of

$$C_{\min}=\sqrt{\frac{\alpha_1 r_1^2+\alpha_2 r_2^2}{R^2}-1}$$

for $\Delta=0$ and a maximum CoV of

$$C_{\max}=\sqrt{\frac{7}{6}\frac{\alpha_1 r_1^2+\alpha_2 r_2^2}{R^2}-1}$$

for $\Delta=1$. Hence, r_1 has to be selected appropriately to provide an adequate range of CoV values $\left[C_{\min};C_{\max}\right]$ to choose from. Note, that after a value for r_1 is determined, all other variables apart from Δ can be computed. Arbitrary CoV values from $\left[0;\infty\right[$ can be generated. For $r_1=R$, we obtain $C_{\min}=0$, and $r_1\to\frac{R}{2}$ leads to $C_{\max}\to\infty$.

Given a target CoV $C\in\left[C_{\min};C_{\max}\right]$ for the bimodal distribution, parameter Δ is computed by

$$\Delta=\sqrt{\frac{6R^2\left(C^2+1\right)}{\alpha_1 r_1^2+\alpha_2 r_2^2}-6}.$$

In summary, the two steps to determine the parameters of the triangular distributions are to find a value for r_1 that provides the appropriate interval $\left[C_{\min};C_{\max}\right]$ and then to fine-tune the CoV by computing Δ with the above formula. All dependent parameters are given explicitly. Thus, no numerical problems are expected for the above formulae.

4 SIMULATION EXPERIMENTS

As test models we used the MIMAC (Measurement and Improvement of MAnufacturing Capacities) test bed datasets 1, 3, 4, 5, 6, 7. Dataset 2 was not used because the simulation package reported problems in the dataset. For further details on the datasets and their download: see `<www.eas.asu.edu/~masmlab>`.

The simulation runs were carried out with Factory Explorer 2.8 from WWK. We simulated 7 years of fab operation with product mixes as given. The first two years were considered as warm-up phase and not taken into account for the statistics. We checked the length of the initial transient both by the cycle time over lot exit time charts and the Schruben test. If there was an indication of initial bias problems the warm-up phase was increased appropriately. The measurement interval was 5 years in all cases. As performance measures we considered the average and the 95%-quantile of the cycle times of all lots.

We simulated factory loads from 70% to 98% of the bottleneck tool group capacity. As dispatching rules we used First In First Out (FIFO) and Critical Ratio (CR) with a target flow factor of 4.0. In the original fab models, all Time To Failure (TTF) and Time To Repair (TTR) distributions were set to exponential. This leads to a coefficient of variation of 1 for TTF and TTR. In our experiments, we replaced the given exponential distributions by Gamma and Weibull distributions or our new bimodal failure model with the same mean values but different Coefficients of Variation. The following Gamma and Weibull distributions were considered:

- CoV = 0.5: less variation than exponential distribution,
- Gamma distribution: shape parameter α = 4,
- Weibull distribution: shape parameter α = 2.10135,
- CoV = 1.0: exponential distribution,
- CoV = 2.0: more variation than exponential distribution,
- Gamma distribution: shape parameter α = 0.25,
- Weibull distribution: shape parameter α = 0.54269.

For the bimodal model we had to replace each of the original unimodal exponential distributions by two triangular distributions running in parallel. Here, it is important that the simulator does not ignore down events that happen during an ongoing down phase of a tool. If this happens the simulator has to prolong the current downtime by the amount of the new downtime.

In Table 1, 2 and 3 we list the average cycle times of all lots for all fab models under a load of 98%. The cycle times are given as multiples of the lots' raw processing times. The tables contain both the FIFO and the CR results. The 95%-quantiles are not given because they show essentially the same behavior as the average cycle time results. As expected, the cycle times increase when the CoV values of the TTF and TTR distributions increase. For the low CoV values the cycle time results match. For the high CoV values, however, the shape of the distributions matters. Whether Gamma, Weibull, or bimodal models lead to higher cycle times depends on the fab model. In all cases, the results for the Exponential distribution are between the low and the high CoV results.

The conclusions from our simulation study are as follows. The CoV values of the TTF and TTR distributions have a considerable impact on the fab performance measures. The effect is less critical for CoV values less than 1 than for CoV values larger than 1. The magnitude of the effects is model dependent. In general, the effects are stronger under FIFO than under CR dispatch regime. In addition to the CoV, the shape of the distribution also matters.

Table 1: Average Cycle Times for CoV=0.5

FIFO	Gamma	Weibull	Bimodal
m1	2.2	2.2	2.4
m3	1.8	1.8	1.8
m4	1.5	1.5	1.6
m5	1.8	1.8	1.8
m6	2.2	2.2	2.2
m7	1.3	1.3	1.3

CR	Gamma	Weibull	Bimodal
m1	3.2	3.2	3.3
m3	2.5	2.5	2.6
m4	1.7	1.7	1.8
m5	2.9	2.8	2.9
m6	2.4	2.4	2.5
m7	1.3	1.3	1.3

Table 2: Average Cycle Times for CoV=1.0

FIFO	Exponential	Bimodal
m1	3.1	3.0
m3	1.8	1.8
m4	1.9	1.9
m5	1.9	1.9
m6	2.6	2.5
m7	1.4	1.5

CR	Exponential	Bimodal
m1	3.8	3.8
m3	2.6	2.6
m4	2.2	2.1
m5	3.0	2.9
m6	2.8	2.8
m7	1.4	1.5

Table 3: Average Cycle Times for CoV=2.0

FIFO	Gamma	Weibull	Bimodal
m1	7.4	6.2	5.4
m3	1.9	1.8	1.9
m4	4.1	4.2	3.0
m5	2.6	3.0	2.7
m6	4.7	3.7	3.4
m7	1.8	1.8	2.0

CR	Gamma	Weibull	Bimodal
m1	4.9	4.8	4.1
m3	2.7	2.7	2.7
m4	4.0	3.1	3.1
m5	3.2	3.3	3.2
m6	5.1	3.5	3.4
m7	2.0	2.2	2.0

Therefore it is questionable to use exponential distributions for modeling TTF and TTR distributions if the CoV value measurements from the fab indicate that this value is not too close to 1. If it is larger than 1 it is recommended to spend some effort on finding the appropriate class of model distributions because the shape of the distribution has a considerable influence on the quality of the results. This does not only hold for the cycle-time estimates but also for the on-time delivery performance estimates as can be concluded from the 95%-quantile results.

5 CONCLUSION

In our research we investigated the appropriateness of non-exponential distributions for modeling machine time to repair (TTR) and time to failure (TTF) in semiconductor wafer fabrication facility models. We examined whether a change in the type of TTR and TTF distributions had an effect on the performance measures, e.g., average cycle times, of a wafer fabrication facility. It turned out that both the variability and the shape of the distribution used for modeling TTR and TTF had a considerable effect on the factory performance estimates. Simplifying assumptions like using exponential distributions for that purpose will cause misleading simulation results.

ACKNOWLEDGMENTS

The author would like to thank Torsten Schiele for his valuable simulation and data analysis efforts.

REFERENCES

Leemis, L. 2001. Input Modeling Techniques for Discrete-event Simulation. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B.A. Peters, J.S. Smith, D.J. Medeiros, and M.W. Rohrer, 62-73. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Schömig, A. 1999. On the Corrupting Influence of Variability in Semiconductor Manufacturing. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans, 837-842. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Schömig, A., and J.W. Fowler. 2000. Modeling Semiconductor Manufacturing Operations. In *The New Simulation in Production and Logistics*, ed. K. Mertins and M. Rabe, 56-64. Berlin.

Schömig, A., and O. Rose. 2003. On the Suitability of the Weibull Distribution for the Approximation of Machine Failures. In *Proceedings of the 2003 Industrial Engineering Research Conference*. Portland, Oregon.

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.

AUTHOR BIOGRAPHY

OLIVER ROSE is Assistant Professor in the Department of Computer Science at the University of Würzburg, Germany. He received an M.S. degree in applied mathematics and a Ph.D. degree in computer science from the same university. He has a strong background in the modeling and performance evaluation of high-speed communication networks. Currently, his research focuses on the analysis of semiconductor and car manufacturing facilities. He is a member of IEEE, ASIM, and SCS. His web address is <www3.informatik.uni-wuerzburg.de/~rose>.