Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

A DISCRETE EVENT SIMULATION MODEL SIMPLIFICATION TECHNIQUE

Rachel T. Johnson
John W. Fowler
Gerald T. Mackulak

Industrial Engineering Dept.
Arizona State University
PO Box 875906
Tempe AZ 85287-5906, U.S.A.

ABSTRACT

Cycle Time – Throughput curves (CT-TH), which plot the average cycle time versus start rate for a given product mix, are often used to support decisions made in manufacturing settings, such as the impact of proposed changes in start rate on mean cycle time. Discrete event simulation is often used to generate estimations of cycle time at a significant number of traffic intensities (start rates). However, simulation often requires long run lengths and extensive output analysis. In most manufacturing environments, the time and/or budget available for such simulations is limited. As demands for faster and more accurate results are required, alternative approaches to improving simulation efficiency must be investigated. This research seeks to develop a procedure for simplifying a detailed model into a fast (abstract) simulation model that achieves a statistically indistinguishable level of accuracy and precision. This technique has particular application in the simulation of semiconductor manufacturing facilities.

1 INTRODUCTION

Discrete event simulation models of semiconductor-manufacturing facilities have proven to be an effective and efficient aid for factory management. Simulation models provide valuable statistical estimates of manufacturing performance measures that support factory decisions regarding many issues such as capacity planning, scheduling, etc. While the power of modern computer packages has greatly risen in recent years, the execution time required to obtain accurate and precise estimates of the statistical outputs from the simulation often requires a large amount of computer execution time.

An example of the number of runs required to obtain a desired confidence interval half-width around an estimated mean cycle-time from a simulation is demonstrated in Figure 1 below.

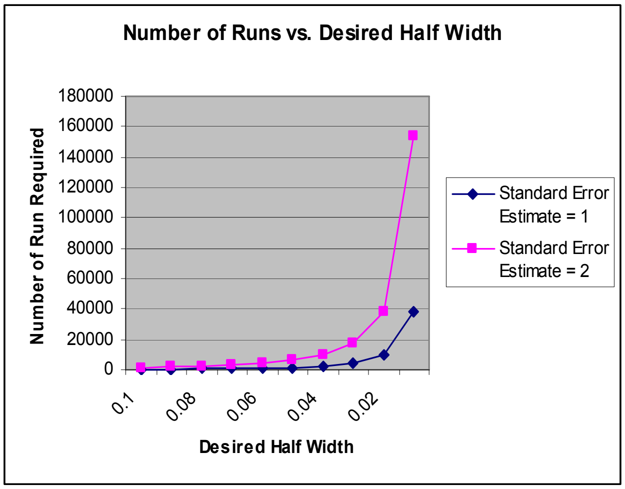

Figure 1: Replications Needed to Obtain a Desired Confidence Interval Half-width of a Mean Cycle-time Estimate

The values in Figure 1 were calculated using equation (1) found in (Law and Kelton 1991). The n value in equation (1)

$$n \geq \left(\frac{t_{1-\alpha/2,n-1} S_{\circ}}{\varepsilon} \right)^2 \qquad (1)$$

represents the amount of replications needed to obtain a given confidence interval half-width, ε. S_o is standard deviation of the mean response estimates calculated from a pilot run of the simulation and $t_{1-\alpha/2,n-1}$ is the Student's t distribution quantile. The estimates of standard deviation in Figure 1 were chosen to be a value of 1.0 and 2.0 for demonstration purposes. One can see that as the desired confidence interval half width decreases – signifying the analyst need for a more precise estimator – the number of runs required to achieve that level of precision increases dramatically. This number of runs can be directly related to the amount of computer effort needed. Often simulations of a manufacturing facility are quite large and the amount of computer time needed to generate an estimator from a single replication is excessive. If that time can be dramatically reduced, the simulation can yield results in a more timely fashion and more replications can be performed.

As a result of the issue concerning model execution time for complex semiconductor manufacturing facility models, several modeling simplification techniques have been proposed. Rose (1998) demonstrates how simple simulation models can be successfully used to explain the behavior of wafer fabs and Rose (1999) demonstrates simple models ability to predict performance measures such as product cycle time. Brooks and Tobias (2000) lay out an eight stage procedure for doing modeling reductions, which result in a simple version of a manufacturing model that is analytically feasible for averages of performance measures. Hung and Leachman (1999) also show that accurate estimates of total cycle time and equipment utilization may be obtained using reduced fabrication simulation models that replace operations at low-utilization workstations with fixed time lags. Peikert et al. (1998) discusses a methodology for quickly investigating problem areas in semiconductor wafer fabrication factories by creating a model for the production area of interest only (as opposed to a model of the complete factory). Thomas and Charpentier (2005) build a simplified simulation model from the bottom up based on a reduced manufacturing routing.

All of the simulation reduction techniques presented vary slightly in approach, but all share the common goal of minimizing the simulation execution time to obtain accurate and precise results (minimizing bias and variance) that are not statistically different from results that would be obtained by a completely detailed model. The most common simplification technique among the papers listed is to retain only the most highly utilized workstations, the bottlenecks, while replacing other workstations with constant delays. Rose (2000) and Hung and Leachman (1999) demonstrate that simple models which use a delay time described by a distribution can fail to describe the detailed model in an accurate way due to lot overtaking (passing). Therefore, it is best to replace removed machine workstations in the simulation model with a constant delay.

While these techniques have proven the ability to provide matching results to a detailed model, and hence (hopefully) the system, as long as proper verification and validation techniques have been applied, none of the aforementioned techniques allows an analytical comparison between the abstract model being created and the detailed model at given points during the model abstraction process. This paper presents a method of sequentially identifying and removing pieces of the model that are "unimportant" to the estimation of the selected performance parameters. This technique illustrates the creation of an abstract model through sequential experiments and demonstrates the models validity by comparing the correlation between the results found in the detailed model and the abstract model. Section 2 presents a high level methodology used to create the abstract model. Section 3 presents an application of the methodology. Section 4 summarizes the findings and presents plans for future investigation.

2 METHODOLOGY

In order to exploit the ability of a simulation to produce performance measure estimates in an economical and efficient way, we propose a technique that allows the identification of model parts that can be replaced by a delay, so as to reduce the model execution time to acquire results without altering the performance of the simulation. To identify these model parts, we show how workstations can be sequentially removed from a simulation model by studying the sample correlation coefficient between the two models (abstract and detailed).

Several assumptions are needed before the model simplification can take place. The first assumption is that the analyst has access to an already built detailed model of the system. The second assumption is that the model has been validated and verified to adequately match the performance of the existing system under study. Finally, the product mix within the model is assumed to be fixed. The steps of the model simplification technique are as follows:

1. Run the detailed model of the system and obtain information regarding the average cycle time a product spends at each workstation in its route and the utilization of the workstations within the model
2. Create a list of machines ordered from the most highly utilized machine to the machine with the lowest utilization
3. Create an abstract model by replacing the bottom X (to be determined by analyst) machines on the list (those with the lowest utilization) with a constant delay that is the sum of the average processing time on the machine and the average time a product spends in the queue at that station
4. Using the same common random numbers employed in the detailed model, run the abstract model with the replaced machines and obtain statistics on average cycle time for the products
5. Measure the sample correlation coefficient found in equation 2

$$\frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{2}$$

through comparison of the average cycle time of a product within a replication of the abstract model (Y) to the average cycle time of a product within a replication of the detailed model (X). (The amount of replications used for comparison should be determined by the analyst, but is recommended to not be lower than 10 replications)

6. If the sample correlation coefficient is above 0.6, which is regarded as highly correlated, continue and return to step number 3. Otherwise, add the last set of removed machines back into the model and stop.

3 SIMULATION MODELS

This section demonstrates an application of the methodology presented in the previous section. The models of interest are a tandem-10 M/M/1 model and a tandem-10 M/M/10 model. Both of these models allow for closed form theoretical calculations of performance measures which can be used to validate the detailed and abstract models abilities to predict the true measures of interest, such as average cycle time. Standard techniques were used for determining a proper warm-up period so data could be truncated for the removal of initial condition bias (Law and Kelton 1991). Additionally, common random numbers were applied so that replications of the abstract and detailed models could be directly compared against each other.

The first model of interest, the tandem-10 M/M/1 model, consists of ten M/M/1 queues in series. The model includes one product that visits each machine once, starting at machine station one and ending at machine station 10 (forward flow only). Machine 4 is the bottleneck machine and has an exponentially distributed processing time with a mean value of 1 time unit. All other machines have exponential service times with a mean processing time of 0.6 time units. This was done so that throughput rate (or traffic intensity) could be equal to the arrival rate. For this study, the model abstraction was done by sequentially removing one machine at a time from the model. This equates to choosing X to equal one in step 3 of the methodology section. It should be noted that removing one machine workstation at a time is generally not recommended because of the large amount of time that would be required to sequentially remove workstations when the model includes hundreds of workstations, such as in a semiconductor manufacturing model. Figure 2, illustrates a comparison of the sample correlation of the abstract model to the detailed model for all nine levels of abstraction in the tandem-10 M/M/1 case. The model number on the X-axis corresponds to the number of machine workstations removed from the model. Four different throughput levels are shown. Ten replications were run for the detailed model and each abstract model.

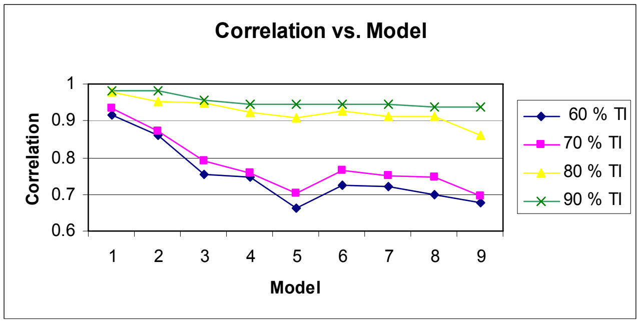

Figure 2: Average Correlation between the Abstract and Detailed Model at 9 Levels of Abstraction

As seen in Figure 2, the correlation coefficient degrades slowly as machines are replaced. Model nine on the chart corresponds to the model that has nine machine workstations replaced by delays and only the bottleneck machine is retained in the model. This model is still highly correlated with the detailed model and was proven to provide accurate and precise estimates of mean cycle time, by comparison to theoretical values. Table 1 below demonstrates that the theoretical values and the mean estimated cycle time of the abstract model both fall in the confidence interval produced by the detailed model.

Table 1: 99% Lower and Upper Bound Confidence Limits for the Tandem 10 - M/M/1 Detailed Simulation Model (Based on 10 Replications) for Each of Four Different Throughput Levels

Throughput	Detailed mean CT	LB CI	UB CI	Theoretical Value	Abstract mean CT
60%	10.90	10.86	10.95	10.94	10.91
70%	12.59	12.52	12.66	12.64	12.59
80%	15.34	15.20	15.48	15.38	15.34
90%	21.64	21.16	22.12	21.74	21.64

Similar results to these were found from the tandem-10 M/M/10 model. This model was identical to the Tandem-10 M/M/1 model on each machine station, but each machine workstation contains ten servers instead of one. Arrival rates were adjusted to reflect this change. Table 1 shows numerically what Figure 2 demonstrated (for the tandem-10 M/M/1 case). but additionally demonstrates what happens when the model abstraction leaves out an important piece of the model.

Table 2: The Correlations between the Detailed and Abstract Models in the Tandem-10 M/M/10 Case. The W*1 and W*9 Models Correspond to Models that Have Retained 1 and 9 Machine Workstations, Respectively, but the Bottleneck Machine is Not Among the Retained Stations.

Model	# of Machine Stations	Correlations			
		60 % TI	70 % TI	80 % TI	90 % TI
1	10	-------------	-------------	-------------	-------------
2	9	0.95	0.95	0.97	1.00
3	8	0.95	0.92	0.95	0.99
4	7	0.81	0.74	0.85	0.99
5	6	0.74	0.67	0.83	0.99
6	5	0.69	0.68	0.84	0.99
7	4	0.54	0.64	0.84	0.99
8	3	0.58	0.63	0.81	0.99
9	2	0.63	0.70	0.86	0.99
10	1	0.54	0.67	0.85	0.99
W*1	1	0.30	-0.14	-0.39	-0.47
W*9	9	0.54	0.24	-0.03	-0.25

One point of interest is to observe what happens to the correlation levels when the bottleneck machine is removed and a machine workstation with a lower utilization is retained. Two models were created for analyzing this scenario. The two models correspond to W*1 and W*9 in the last two rows of table 2. W*1 is a model containing only one machine work station (comparable to model 10), but the one retained machine is not the bottleneck machine. W* 9 is a model containing 9 machine work stations (comparable to model 2), but again omitting the bottleneck machine from the model. From the last two rows in Table 2, one can see that when a significant piece of the model is removed, the correlation values degrade significantly. The average correlation across the four throughput rates for Model 10 is approximately .76 where as the average correlation across the four throughput rates for W*1 is approximately -.18. This signifies a considerable deterioration. Similar results are seen when comparing Model 2 to W*9.

While the correlation coefficients were seen to significantly drop, it is noteworthy to mention that the W* 1 and W* 9 models still produced estimates of mean cycle time that fell within the confidence limits produced by the detailed model. However, if the analyst was to look at the autocorrelation between the detailed model and incorrect abstract models, considerable differences would be found. Figures 3, 4, and 5 show the autocorrelation graphs of the detailed model, abstract model number 10, and abstract model W*1. The detailed and abstract model number 10 models show similar autocorrelation graphs, where as the W*1 model significantly deviates in structure from the other two. This is important because the autocorrelation of the output from the models is related to the distribution of the output data. Different distributions would lead to significantly different results when comparing percentiles and quantiles, which are often used in the manufacturing setting.

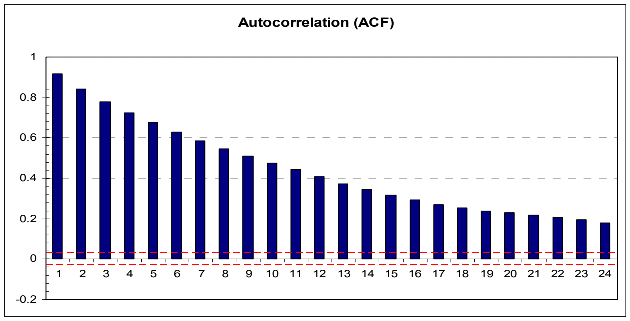

Figure 3: Autocorrelation Graph of the Detailed Model, at a 70% Traffic Intensity, with Results Based on a Single Replication of Output Data

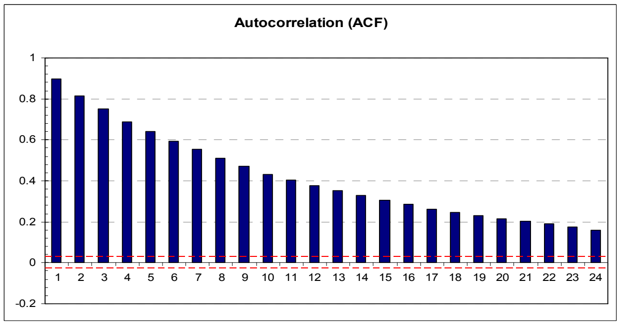

Figure 4: Autocorrelation Graph of the Abstract Model Number 10, at 70% Traffic Intensity, with Results Based on a Single Replication of Output Data

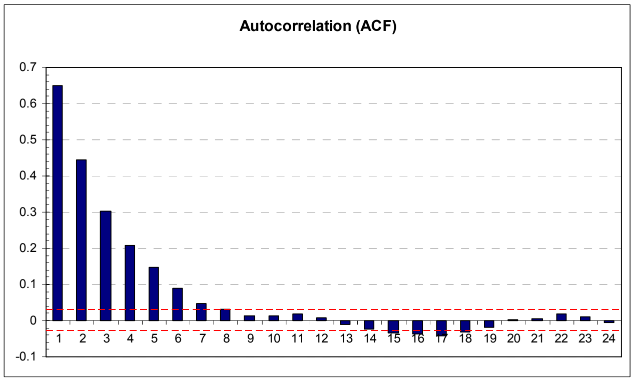

Figure 5: Autocorrelation Graph of the W*1 Model, at 70% Traffic Intensity, with Results Based on a Single Replication of Output Data

The autocorrelation graphs were all created by using a single replication of output data. The autocorrelations measure the correlation between observations and a lag of 1 to 24 was used for the graph. The correlation is labeled on the Y axis, while the lag is labeled on the X axis. The graphs were created by using a freeware package found at: <www.web-reg.de>, created by Kurt Annen.

4 CONCLUSIONS AND FUTURE WORK

A method for creating an abstract simulation model from a highly detailed one by sequentially replacing pieces of the model with delays and checking to make sure the two models were correlated during each step of the abstraction was presented. Initial model testing done on two tandem-10 M/M/c queues was presented and demonstrated promising results.

It was shown that the abstract model demonstrated a high correlation value to the detailed model in all cases where the bottleneck machine remained in the model. Also, the abstract models were able to match mean cycle times of the detailed model and the output observation were shown to have similar autocorrelation functions when the most highly utilized machines workstations were retained in the abstract model.

Future work will include testing the algorithm on a real world semiconductor manufacturing simulation model. Several test beds for this type of testing exist on the website provided by the Modeling and Analysis of Semiconductor Manufacturing (MASM) lab at Arizona State University <http://www.eas.asu.edu/~masmlab/ftp.htm>.

REFERENCES

Brooks, R. J. and Tobias, A. M. 2000. Simplification in the simulation of manufacturing systems. *International Journal of Production Research*, Vol. 38, No. 5, 1009-1027.

Hung, Y. F. and Leachman, R.C. 1999. Reduced simulation models of wafer fabrication facilities. *International Journal of Production Research*, Vol. 37, No. 12, 2685-2701.

Law, A. M. and Kelton, W. D. 1991. *Simulation Modeling & Analysis* (3rd ed.) New York: McGraw-Hill.

Peikert, A., Thoma, J. and Brown, S. 1998. A rapid modeling technique for measurable improvements in factory performance. *Proceedings of the 1998 Winter Simulation Conference*, 1011-1015.

Rose, O. 1998. WIP evolution of a semiconductor factory after a bottleneck workcenter breakdown. *Proceeding of the 1998 Winter Simulation Conference*, 997 – 1003.

Rose, O. 1999. Estimation of the cycle time distribution of a wafer fab by a simple simulation model. *Proceedings of the SMOMS '99* (1999 WMC), 133 – 138.

Rose, O. 2000. Why do simple wafer fab models fail in certain scenarios? *Proceedings of the 2000 Winter Simulation Conference*, 1481 – 1490.

Thomas, A. and Charpentier, P. 2005. Reducing simulation models for scheduling manufacturing facilities. *European Journal of Operational Research*, 161, 111-125.

AUTHOR BIOGRAPHIES

RACHEL T. JOHNSON is a graduate student in the Industrial Engineering department at Arizona State University. Her research interest is in discrete event simulation methodologies. She is a member of INFORMS and served as the INFORMS student chapter treasurer for the school year 2004-2005. She received her B.S. in Industrial Engineering from Northwestern University. She was awarded the SRC/Intel Fellowship in the fall of 2004 for the duration of her Masters program.

JOHN W. FOWLER is a Professor of Industrial Engineering at Arizona State University (ASU). Prior to his current position, he was a Senior Member of Technical Staff in the Modeling, CAD, and Statistical Methods Division of SEMATECH. He spent the last year and a half of his doctoral studies as an Intern at Advanced Micro Devices. His research interests include modeling, analysis, and control of manufacturing (especially semiconductor) systems. He is the Co-Director of the Modeling and Analysis of Semiconductor Manufacturing Laboratory at ASU. The lab has had research contracts with NSF, SRC, International SEMATECH, Intel, Motorola, Infineon Technologies, ST Microelectronics, and Tefen, Ltd. Dr. Fowler is a member of ASEE, IIE, INFORMS, POMS, and SCS. He is an Area Editor *for SIMULATION: Transactions of the Society for Modeling and Simulation International* and an Associate Editor of *IEEE Transactions on Electronics Packaging Manufacturing*.

GERALD T. MACKULAK is an Associate Professor in the Department of Industrial Engineering at Arizona State University. He is a graduate of Purdue University receiving his B.Sc., M.Sc., and Ph.D. degrees in the area of Industrial Engineering. His primary area of research is in extending the methodology of simulation to a broader user base. For the past several years he has been concentrating on simulation applied to semiconductor manufacturing.