Proceedings of the 2007 INFORMS Simulation Society Research Workshop.

MULTI-PRODUCT CYCLE TIME AND THROUGHPUT EVALUATION VIA *SIMULATION ON DEMAND*

John W. Fowler
Gerald T. Mackulak

Department of Industrial Engineering
Arizona State University
Tempe, AZ, 85287-5906, U.S.A.

Barry L. Nelson
Bruce Ankenman

Dept. of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL, 60208-3119, U.S.A.

ABSTRACT

In this paper, we will discuss our efforts to create the next generation of semiconductor factory simulation tools, which we call complete response-surface mapping (cRSM). More specifically, we will describe the basic research and software development necessary to produce the capability to provide simulation results *on demand* for cycle-time measures as a function of throughput and product mix.

1 INTRODUCTION

Many man-hours are invested in developing and exercising simulation models of wafer fabs, models that include critical details that are difficult or impossible to incorporate into simple load calculations or queueing approximations. Unfortunately, simulation models can be clumsy tools for planning or decision making because even a few minutes per simulation run (which is optimistic) is too slow to allow what-if analysis in real time. Even optimization via simulation (where some combination of simulation outputs is maximized or minimized) has drawbacks since an objective function must be specified and this hinders the decision maker's ability to consider trade offs that are not easily quantified.

We are investigating the creation of the next generation of simulation tools for decision support in semiconductor manufacturing, which we call complete response-surface mapping (cRSM). cRSM exploits the availability of large quantities of idle computer resources, while recognizing the scarcity of decision-maker time. cRSM combines computing horsepower, adaptive statistical methods and queueing theory to allow a simulation to be used for planning and decision making in a much different way than before. cRSM represents a bridge between the flexibility of simulation and the insight provided by an analytical queueing model by delivering simulation results *on demand*.

More specifically, we are performing the basic research and software development to produce a cRSM tool that provides simulation results on demand for cycle-time measures as a function of throughput and product mix in semiconductor manufacturing. Given a simulation model of a wafer fab and minimal information on the controllable parameters, cRSM runs an automated sequence of experiments to generate a *model structure* (MS) that represents the first four moments (equivalently, mean, variance, skewness and kurtosis) of product cycle time as a function of product mix and throughput. These experiments could use idle computer resources, exploit multiple processors if they are available, and execute without human intervention. The MS is the input to a simulation-on-demand *query engine* (QE) that allows the decision maker to investigate options and trade offs on demand *without running additional simulations*. Any questions that can be answered through combinations of the mean, standard deviation and percentiles of the cycle time as a function of throughput and product mix are supported, with results delivered as numerical and graphical displays.

To be more precise, let $l_1, l_2, \ldots, l_k$ be the throughputs (release rates) of k products into the factory simulation. We denote the steady-state cycle time of the j th product by

$$C_j = C_j(l_1, \ldots, l_k) = C_j(l, a_1, \ldots, a_k)$$

where l is the factory throughput, and a_j is the fraction of the throughput that will be product type j ($\sum_{j=1}^{k} a_j = 1$). cRSM will produce a MS that approximates the moments of C_j as the decision variables $l, a_1, \ldots, a_k$ vary over their feasible values. This allows the decision maker to answer questions such as:

1. What is the weighted cycle time of the factory at a particular throughput and product mix?
2. What is the 80th percentile of cycle time for products at a particular throughput and product mix?
3. What are the feasible values of throughput and product mix $l, a_1, \ldots, a_k$ such that average cycle-time constraints $E[C_j] \le c_j, j = 1,2,\ldots,k$ are met?

4. What is the impact on the cycle times of products ***1,2,…, k-1*** of increasing the throughput of product ***k*** to meet increased demand?
5. What product mix maximizes revenue while keeping cycle times below required limits?

The cRSM that we are developing builds an MS for a factory simulation in which only the throughput and product mix can be altered; however, multiple instances of a cRSM can be used for capacity planning objectives. By letting cRSM build an MS for factory simulations with different levels of capacity, cRSM facilitates capacity planning and expansion analysis that takes cycle time into account.

The emphasis of this approach is different from much simulation research: Our focus is on the efficiency of obtaining useful simulation results, rather than on the efficiency of the simulation run itself. cRSM assumes that the user is willing to run a substantial number of simulations to build the MS, although "substantial" still means orders of magnitude less time than was required to build the simulation model. We will design cRSM to make these runs efficiently, but the real savings from cRSM are most apparent *after* the MS is available, when a decision maker can use the QE to quickly and easily answer a variety of questions on demand, without rerunning the simulation or even knowing that a simulation exists. Our goal is to get more value out of the simulation, via deeper insight, more complete exploration and timely responses, than is currently possible with either simulation or analytical models.

2 PRIOR WORK

In our previous NSF-sponsored research we developed efficient tools for accurate and precise estimation of the mean, standard deviation and percentiles of cycle time as a function of the overall factory throughput for a fixed product mix. Central to this work was developing flexible families of models to represent the first four moments of steady-state cycle time as a function of throughput; percentiles of cycle time then come from a four-moment approximation. To design simulation experiments to fit these models, we also needed models for the variability of the moment estimators themselves, because the moments, and the variance of their estimators, explode as the throughput approaches factory capacity. The form of our models was motivated by heavy-traffic queueing analysis (e.g., Whitt 1989), but our research showed that generalizations of these simple models were essential when the simulation included tool failures, different product flows and priority schemes found in semiconductor manufacturing (Allen 2003, Johnson 2003).

To define these models, let x represent the fraction of system capacity in use when the factory throughput capability is l (this allows the maximum throughput to be standardized as 1). We developed techniques to use wafer fab simulation outputs to fit the following model for the ***m***th moment of cycle time:

$$E[C^m] = \frac{\sum_{i=0}^{t} a_i x^i}{(1-x)^p} \qquad (1.1)$$

We also needed a model for the variance of the estimator of the ***m***th moment, specifically

$$Var[\bar{C}^m] = \frac{\sum_{i=0}^{u} b_i x^i}{(1-x)^{2q}} \qquad (1.2)$$

For simple networks of first-in-first-out queues, heavy traffic analysis suggests that $p=m$, $q=2m+2$; however, we showed that this is not always the case for the queueing networks that are typical of semiconductor manufacturing facilities. With both p and q unknown, and up to eight models to be fit simultaneously ((1.1) and (1.2) for each of the first four moments), we developed techniques to efficiently and effectively design the simulation experiment and fit the models, and showed that these models give remarkably accurate predictions of the mean, standard deviation and percentiles of cycle time (McNeill *et al.* 2003ab, Mackulak *et al.* 2004, Park, *et al.* 2002, Yang *et al.* 2004). Our experience building CT-TH models for a fixed product mix demonstrates that simulation output data can be used to fit accurate, and easily manipulated, models of the form (1.1) for the factory as a whole. Once the models are available they can be used to quickly and interactively evaluate cycle time-throughput scenarios on demand in the same way we use a queueing model. Our approach is to take these ideas to the next level in two important ways: (1) To allow product mix, as well as throughput, to be varied; and (2) to develop a simulation-on-demand QE that uses these models for decision support.

3 MODELS FOR PRODUCT MIX

We believe that the best approach for incorporating product mix into the CT-TH analysis is to leverage, as much as possible, our expertise in developing CT-TH models with a fixed product mix. For a given product mix, we have developed procedures to fit mean cycle time curves as a function of the throughput, and then to derive cycle-time percentile curves from these moment models. The procedures make efficient use of the simulation runs, diagnose and correct for lack of fit, and can be completely automated. Unfortunately, our investigation of analytically tractable queueing network models convinces us that extending the moment model (1.1) to include product mix as an independent variable (as in Lamb and Cheng 2002) is unlikely to be successful because the correct form of the model depends on specifics of the network topology of the factory, something we do not think the user of cRSM should have

to figure out. Instead, we will use simulation to fit CT-TH models for a carefully selected range of product mixes and then interpolate among these models to derive cycle-time measures at product mixes that we did not simulate. In a rough sense, we are looking for a set of basis functions that span the cycle time for the product mix space of interest. Since we already have the capability to fit CT-TH slices of this surface (i.e., curves for a fixed product mix), the focus of the research is designing the simulation experiment, interpolating the curves, and verifying the accuracy of the results. We next discuss each of these issues in turn

3.1 Design of the cRSM Experiment

Our previous research has provided efficient and effective experiment design strategies when the product mix is fixed. In this context a "design" corresponds to settings of the standardized throughput at which to make simulation runs, and an allocation of simulation effort to each design point (Park *et al*. 2002, Yang *et al*. 2004). For cRSM, the design also includes the product mix settings at which we fit the models. There are a number of research challenges to address:

1. The design space is no longer simple as the stability requirement (throughput must be less than capacity) for different product mixes further complicates the design problem. The work center, machine group or station that first reaches capacity *depends on the product mix* $(a_1, \ldots a_k)$. Figure 1a shows the CT-TH curves for Product 1 in a two product, multi-station system when the mix of Product 1 changes.
2. The experiment design must fill the product mix space in a way that facilitates interpolation at product mixes not simulated. Thus, a good design might need to include settings of product mix that cause each work center, machine group or station that *could* define the fab capacity to actually define it.

3.2 Interpolation

A second research challenge is interpolating among the fitted CT-TH curves when we encounter a new product mix. For instance, in the example shown in Figure 1 the curve for Product 1 at 50% of the mix would be some interpolation of the four fitted curves. What kind of interpolation will work best?

It is plausible that CT-TH curves derived at a base collection of product-mix settings can be used to infer the entire CT-TH surface, but we cannot expect the interpolation to be so easy in a practical situation. The following is a more realistic approach: Let $\mathbf{a} = (a_1, \ldots a_k)$ denote a vector of mix parameters, and let $\mathbf{A}$ denote the collection of product mixes at which we run simulations and fit models. Then for a new mix $\mathbf{a}'$ that we did not simulate, the interpolation might take the form

$$\hat{C}_j(l, \mathbf{a}') = \sum_{\mathbf{a} \in ?} D(\mathbf{a}, \mathbf{a}') \hat{C}_j(l, \mathbf{a}) \qquad (1.3)$$

where D is a measure of the distance between the mix of a fitted curve and the desired curve, and $\hat{C}_j(l, \mathbf{a})$ is the fitted curve at mix $\mathbf{a}$. Figure 1 illustrates how this might be done to determine Product 1's cycle time when it is 50% of the mix, and the fab throughput is 4. In this case the values of the four fitted curves at 20%, 40%, 60% and 80% Product 1 provide four values for a quadratic interpolation at 50% Product 1 (right figure). The interpolated value is 0.1545 days, while the true mean cycle time is 0.1549 days. Although the quadratic distance measure D works well in this example, the key research question will be the choice of distance measure for more complex problems.

3.3 Accuracy

Since we require the experiment design, simulation and fitting process (in other words, the construction of the MS) to be completely automated, we need a way to determine when the MS is complete. We will again leverage our ability to construct accurate and precise CT-TH curves for a fixed product mix, implying that we have confidence that each individual curve in our "basis" is valid. To determine when the basis is complete, we propose using a cross-validation approach. When we have fit g curves, we can test these curves for adequacy by dropping one curve at a time and measuring the ability of the remaining $g-1$ curves to approximate it via (1.5). When all curves can be accurately approximated by interpolating the other $g-1$ curves then the MS is complete. Based on our experience fitting CT-TH curves, maximum relative error along the curve is a good choice for measuring approximation accuracy.

4 SOFTWARE TOOLS

To support the development and use of cRSM, two software tools must be built. The first, the cRSM Generator, is for use in creating the Model Structure (MS) and will control which combinations of throughput rates and product mixes are simulated. The second is the simulation-on-demand Query Engine (QE), which is for use after the MS has been generated. The QE answers questions, such as those outlined above, given a collection of nonlinear response surface models for each of the first four moments at various product mixes. Results will be provided *on demand*, without requiring any additional simulation effort.

4.1 cRSM Generator

A critical step in creating the cRSM that must occur before the MS is generated is performing simulation runs at a variety of product mix/throughput combinations. Since it is impossible to perform simulation runs at all possible combinations, the cRSM Generator determines what design points (throughputs and product mixes) to run, how much simulation effort to allocate at each design point and automatically executes the runs. It requires as input the product routings, the processing rates of each product on each machine group, and the number of machines in each machine group. Our prior work focused on a single product mix, which essentially reduces the problem to a single slice through the response surface. Adding the additional variable of product mix makes the problem significantly more difficult, as it dramatically increases the number of candidate design points at which to simulate. Further, it is also important to determine the relative importance of each design point for developing the MS. Both tasks are accomplished by the cRSM Generator. Finally, since the results of all simulation runs will be stored in a database, the cRSM generator will have the capability to reuse existing simulation runs. For example, if answering question posed to the Query Engine requires additional precision beyond that attained by the initial cRSM runs, then the initial results will be retrievable so that duplicate effort is not required to obtain the new, higher-precision results

4.2 Query Engine

Once the MS has been generated, nonlinear response surface models exist for each of the first four moments of cycle time. These models can be used to answer a variety of interesting questions about the system, and the simulation-on-demand QE will provide the decision support mechanism by which these questions are asked and answered. Specifically, the QE software tool will consist of a front-end user interface, which accepts inputs from the user and displays the outputs. Additionally, to answer several of the questions we anticipate decision makers asking, the QE will need to contain a response-surface search algorithm that efficiently and effectively searches the CT-TH surfaces to find settings that yield specified cycle-time performance, and locate optimal or near-optimal solutions. There are two types of questions that will typically be handled through the QE: those about the mean cycle time of the system and those about cycle-time percentiles. Specific solution approaches for some sample questions are given below.

Question 1: What are the mean (or P-percentile) product cycle times for a given fab throughput and product mix or vector of start rates?

In this case, the user is interested in obtaining an estimate of the mean *(or P-percentile)* cycle time for each product when the start rates of all products are already known. The input can be given as an overall factory throughput and product mix or as a vector of start rates for each product type. To answer this question using the QE, the appropriate coordinates on the mean response surface model (the independent axes represent the start rates for each of the products) simply need to be identified and the associated response interpolated for each product. When the user is interested in obtaining a vector of estimates of a particular cycle time percentile, (i.e., the vector of the 95th cycle-time percentiles) for a given set of start rates, the answer relies on our previous work in percentile estimation using the Cornish-Fisher expansion, which was found to effectively estimate percentiles from any sample distribution, given a percentile from the standard normal distribution and estimates of the sample distribution's first four moments. The only required user inputs are the desired percentile and the product mix. Further details on the Cornish-Fisher expansion as it applies to cycle-time percentile estimation can be found in the McNeill, *et al.* (2003ab).

Question 2: What product mix is most profitable that achieves given mean (percentile) cycle time targets?

The user is now interested in determining the product mix (or, equivalently, vector of start rates) that will maximize profit for the system, while still obtaining no more than a maximum mean (or percentile) cycle-time value for each product. To answer this question, the user must supply the following inputs: the vector of profits for each unit of each product produced, the cycle time requirements for each product; the minimum start rate for each product (based on production requirements) the maximum start rate for each product (based on product demand). Obtaining a solution to this question will be significantly more difficult than the previous question, as there may be an infinite number of product mixes that will meet the cycle-time requirements. Therefore, we must find an efficient way to navigate the solution space towards the optimum, or at least toward a solution with a high total profit. We expect to use nonlinear programming techniques, such as gradient search that is specialized to exploit the fact that these nonlinear functions have a known general form (1.1) and certain properties (e.g., monotonically increasing as throughput increases).

If the cycle time targets are on percentiles, an additional step, evaluating the Cornish-Fisher (C-F) expansion, will be required. Once a possible solution is identified, each of the first four sample moments must be estimated for each product using the CT-TH surface models. These values will then be plugged into the C-F expansion to determine if the percentile estimate for each product meets the constraint. If the constraint vector is met, the profit for this solution is stored, and the solution space search continues. If the constraint is not met, the solution is infeasible, and the objective function value need not be calculated. Alternatively, we will investigate whether it is more

efficient to build an entire response surface of the Cornish-Fisher expansion for a given percentile, which could be searched directly, rather than evaluating the expansion independently for each throughput and product mix. Clearly, Question 2 is more difficult than Question 1 because it involves a complex feasible region. However, answering this type of question provides great benefit to the decision maker and searching the solution space will take significantly less time than running additional simulation models or trying to do optimization via simulation.

Figure 2 shows the relationship between the user inputs, the MS inputs, and the outputs of the QE for the questions discussed. Neither the list of questions nor the figure is intended to be exhaustive. Rather, they are intended to provide an example of the types of questions that the QE will be able to answer.

REFERENCES

Allen, C. 2003. "The Impact of Network Topology on Rational-Function Models of the Cycle Time-Throughput Curve," Honors Thesis, Department of Industrial Engineering & Management Sciences, Northwestern University.

Johnson, R. 2003. "Non-Linear Regression Fits for Cycle Time vs. Throughput Curves using Two Data Sets from Actual Semiconductor Manufacturing Facilities," Research Report, Dept of Industrial Engineering & Management Sciences, Northwestern University.

Lamb, J. and R. Cheng. 2002. Optimal allocation of runs in a simulation metamodel with several independent variables. *Operations Research Letters* **30**, 189-194.

Mackulak, G., Fowler, J., Park, S., McNeill, J.E. 2004. "A Three Phase Simulation Methodology for Generating Accurate and Precise Cycle Time - Throughput Curves", accepted by International Journal of Simulation and Process Modeling, Vol. 1, Nos. 1/2, pp 36-47.

McNeill, J.E., Mackulak, G.T., and Fowler, J.W., (2003a) "Indirect Estimation of Cycle Time Quantiles From Discrete Event Simulation Models Using the Cornish-Fisher Expansion," Proceedings of the 2003 Winter Simulation Conference, S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds., pp 1378-1382.

McNeill, J.E., Mackulak, G.T., Fowler, J.W., and Nelson, B.L. (2003b) "Indirect Cycle Time Quantile Estimation Using the Cornish-Fisher Expansion," ASU Working Paper Series.

Park, S., Fowler, J., Mackulak, G., Keats, J. and Carlyle, W. (2002) "D-optimal sequential experiments for generating a simulation-based cycle time-throughput curve," Operations Research, **50(6)**, pp. 981-990.

Yang, F., Ankenman, B. and Nelson, B.L. 2004. "Generation of Cycle Time-Throughput Curves through Simulation," Working Paper, Dept of Industrial Engineering & Management Sciences, Northwestern University.

Whitt, W. 1989. "Planning Queueing Simulations," Management Science 35, 1341-1366..

AUTHOR BIOGRAPHIES

JOHN W. FOWLER is a Professor of Industrial Engineering at Arizona State University (ASU) and is the Center Director for the Factory Operations Research Center that is jointly funded by International SEMATECH and the Semiconductor Research Corporation. His research interests include modeling, analysis, and control of semiconductor manufacturing systems. 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. He is an IIE Fellow and is on the Winter Simulation Conference Board of Directors. His email address is <john.fowler@asu.edu>.

BARRY L. NELSON is the Krebs Professor of Industrial Engineering and Management Sciences at Northwestern University, and is Director of the Master of Engineering Management Program there. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems. He has published numerous papers and two books. Nelson has served the profession as the Simulation Area Editor of *Operations Research* and President of the INFORMS (then TIMS) College on Simulation. He has held many positions for the Winter Simulation Conference, including Program Chair in 1997 and currently Chair of its Board of Directors. His e-mail address is <nelsonb@northwestern.edu>

GERALD T. MACKULAK is an Associate Professor of Engineering 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 simulation applications within manufacturing with a special focus on automated material handling within semiconductor manufacturing. His email address is <mackulak@asu.edu>.

BRUCE E. ANKENMAN is an Associate Professor in the Department of Industrial Engineering and Management Sciences at the McCormick School of Engineering at Northwestern University. His current research interests include response surface methodology, design of experiments, robust design, experiments involving variance components and dispersion effects, and design for simulation experiments. He is a past chair of the Quality Statistics and Reliability Section of INFORMS, is an Associate Editor for *Naval Research Logistics* and is a Department Editor for *IIE Transactions: Quality and Reliability Engineering*. His e-mail address is ankenman@northwestern.edu.

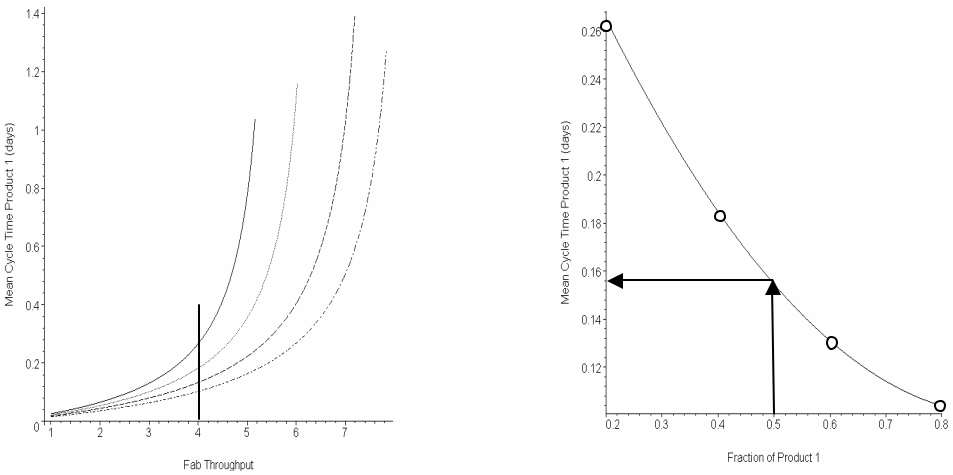

Figure 1: (a) CT-TH curves for one product in a two-product system as the mix changes from (left to right) 20%, 40%, 60% to 80% of Product 1. (b) Interpolation of CT-TH curves with 20%, 40%, 60% and 80% Product 1 to determine cycle time at 50% Product 1.

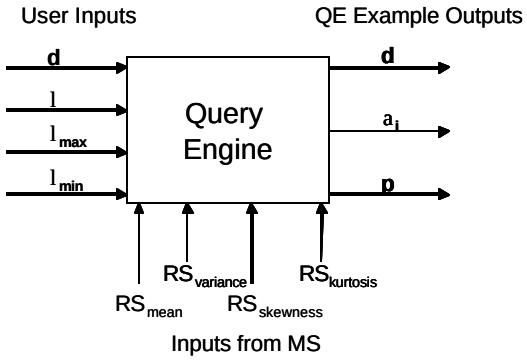

Figure 2: Representation of the QE in terms of potential inputs and outputs